Only Connect?

A Critical Appraisal of Connecting Practices in the Age of Social Media

Digital Methods SUMMER SCHOOL 2016

[reader]

**27 JUNE -
08 JULY
2016**

#dmi16

DIGITAL
METHODS
SUMMER
SCHOOL

Table of Contents

Week 1: Connective action, global diaspora studies and digital methods

Table of Contents

Week 2: Data activism

Beer, D. (2016). How should we do the history of Big Data? @ge B_r_ $ Qmagcrw* 3(1). doi:10.1177/2053951716646135

Couldry, N., & Powell, A. (2014). Big Data from the Bottom Up. @ge B_r_ $ Qmagcrw, 1(2), 1–5.

Dalton, C. & Thatcher, J. (2014). What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'. *Space and Society Open Site.* https://societyandspace.com/material/commentaries/craig-dalton-and-jim-thatcher-what-does-a-critical-data-studies-look-like-and-why-do-we-care-seven-points-for-a-critical-approach-to-big-data/

Powell, A. (2016). Hacking in the public interest: Authority, legitimacy, means, and ends. *New Media & Society*, 18(4), 600–616.

Renzi, A., & Langlois, G. (2015). Data activism. In G. Langlois, J. Redden & G. Elmer (Eds). *Compromised data: From social media to Big Data*. London: Bloomsbury Publishing.

Schrock, A. R. (2016). Civic hacking as data activism and advocacy: A history from publicity to open government data. *New Media & Society*, 18(4), 581–599.

Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, 43(3), 377–391.

Taylor, L. (2015). Towards a contextual and inclusive data studies: A response to Dalton and Thatcher. *Space and Society Open Site*. https://societyandspace.com/material/commentaries/linnet-taylor-towards-a-contextual-and-inclusive-data-studies-a-response-to-dalton-and-thatcher/

W. Lance Bennett & Alexandra Segerberg

THE LOGIC OF CONNECTIVE ACTION
Digital media and the personalization of contentious politics

From the Arab Spring and los indignados in Spain, to Occupy Wall Street (and beyond), large-scale, sustained protests are using digital media in ways that go beyond sending and receiving messages. Some of these action formations contain relatively small roles for formal brick and mortar organizations. Others involve well-established advocacy organizations, in hybrid relations with other organizations, using technologies that enable personalized public engagement. Both stand in contrast to the more familiar organizationally managed and brokered action conventionally associated with social movement and issue advocacy. This article examines the organizational dynamics that emerge when communication becomes a prominent part of organizational structure. It argues that understanding such variations in large-scale action networks requires distinguishing between at least two logics that may be in play: The familiar logic of collective action associated with high levels of organizational resources and the formation of collective identities, and the less familiar logic of connective action based on personalized content sharing across media networks. In the former, introducing digital media do not change the core dynamics of the action. In the case of the latter, they do. Building on these distinctions, the article presents three ideal types of large-scale action networks that are becoming prominent in the contentious politics of the contemporary era.

Keywords collective action; contentious politics; digital media

(*Received 14 November 2011; final version received 22 February 2012*)

With the world economy in crisis, the heads of the 20 leading economies held a series of meetings beginning in fall of 2008 to coordinate financial rescue policies. Wherever the G20 leaders met, whether in Washington, London, St. Andrews, Pittsburgh, Toronto, or Seoul, they were greeted by protests. In London,

Information, Communication & Society Vol. 15, No. 5, June 2012, pp. 739–768
ISSN 1369-118X print/ISSN 1468-4462 online © 2012 Taylor & Francis
http://www.tandfonline.com http://dx.doi.org/10.1080/1369118X.2012.670661

Routledge
Taylor & Francis Group

anti-capitalist, environmental direct activist, and non-governmental organization (NGO)-sponsored actions were coordinated across different days. The largest of these demonstrations was sponsored by a number of prominent NGOs including Oxfam, Friends of the Earth, Save the Children, and World Vision. This loose coalition launched a *Put People First* (PPF) campaign promoting public mobilization against social and environmental harms of 'business as usual' solutions to the financial crisis. The website for the campaign carried the simple statement:

> Even before the banking collapse, the world suffered poverty, inequality and the threat of climate chaos. The world has followed a financial model that has created an economy fuelled by ever-increasing debt, both financial and environmental. Our future depends on creating an economy based on fair distribution of wealth, decent jobs for all and a low carbon future. (Put People First 2009)

The centerpiece of this *PPF* campaign was a march of some 35,000 people through the streets of London a few days ahead of the G20 meeting to give voice and show commitment to the campaign's simple theme.

The London PPF protest drew together a large and diverse protest with the emphasis on personal expression, but it still displayed what Tilly (2004, 2006) termed WUNC: *worthiness* embodied by the endorsements by some 160 prominent civil society organizations and recognition of their demands by various prominent officials; *unity* reflected in the orderliness of the event; *numbers* of participants that made PPF the largest of a series of London G20 protests and the largest demonstration during the string of G20 meetings in different world locations; and *commitment* reflected in the presence of delegations from some 20 different nations who joined local citizens in spending much of the day listening to speakers in Hyde Park or attending religious services sponsored by church-based development organizations.[1] The large volume of generally positive press coverage reflected all of these characteristics, and responses from heads of state to the demonstrators accentuated the worthiness of the event (Bennett & Segerberg 2011).[2]

The protests continued as the G20 in 2010 issued a policy statement making it clear that debt reduction and austerity would be the centerpieces of a political program that could send shocks through economies from the United States and the UK, to Greece, Italy, and Spain, while pushing more decisive action on climate change onto the back burner. Public anger swept cities from Madison to Madrid, as citizens protested that their governments, no matter what their political stripe, offered no alternatives to the economic dictates of a so-called neoliberal economic regime that seemed to operate from corporate and financial power centers beyond popular accountability, and, some argued, even beyond the control of states.

Some of these protests seemed to operate with surprisingly light involvement from conventional organizations. For example, in Spain '*los indignados*' (the indignant ones) mobilized in 2011 under the name of 15M for the date (May 15) of the mass mobilization that involved protests in some 60 cities. One of the most remarkable aspects of this sustained protest organization was its success at keeping political parties, unions, and other powerful political organizations out: indeed, they were targeted as part of the political problem. There were, of course, civil society organizations supporting 15M, but they generally stayed in the background to honor the personalized identity of the movement: the faces and voices of millions of ordinary people displaced by financial and political crises. The most visible organization consisted of the richly layered digital and interpersonal communication networks centering around the media hub of *Democracia real YA!*[3] At the time of this writing, this network included links to over 80 local Spanish city nodes, and a number of international solidarity networks. On the one hand, *Democracia real YA!* seemed to be a website and on the other, it was a densely populated and effective organization. It makes sense to think of the core organization of the *indignados* as both of these and more, revealing the hybrid nature of digitally mediated organization (Chadwick 2011).

Given its seemingly informal organization, the 15M mobilization surprised many observers by sustaining and even building strength over time, using a mix of online media and offline activities that included face-to-face organizing, encampments in city centers, and marches across the country. Throughout, the participants communicated a collective identity of being leaderless, signaling that labor unions, parties, and more radical movement groups should stay at the margins. A survey of 15M protesters by a team of Spanish researchers showed that the relationships between individuals and organizations differed in at least three ways from participants in an array of other more conventional movement protests, including a general strike, a regional protest, and a pro-life demonstration: (1) where strong majorities of participants in other protests recognized the involvement of key organizations with brick and mortar addresses, only 38 per cent of *indignados* did so; (2) only 13 per cent of the organizations cited by 15M participants offered any membership or affiliation possibilities, in contrast to large majorities who listed membership organizations as being important in the other demonstrations; and (3) the mean age range of organizations (such as parties and unions) listed in the comparison protests ranged from 10 to over 40 years, while the organizations cited in association with 15M were, on average, less than 3 years old (Anduiza *et al*. 2011). Despite, or perhaps because of, these interesting organizational differences, the ongoing series of 15M protests attracted participation from somewhere between 6 and 8 million people, a remarkable number in a nation of 40 million (rtve 2011).

Similar to PPF, the *indignados* achieved impressive levels of communication with outside publics both directly, via images and messages spread virally across social networks, and indirectly, when anonymous Twitter streams and YouTube videos were taken up as mainstream press sources. Their actions became daily news fare in Spain and abroad, with the protesters receiving generally positive coverage of their personal messages in local and national news – again defying familiar observations about the difficulty of gaining positive news coverage for collective actions that spill outside the bounds of institutions and take to the streets (Gitlin 1980).[4] In addition to communicating concerns about jobs and the economy, the clear message was that people felt the democratic system had broken to the point that all parties and leaders were under the influence of banks and international financial powers. Despite avoiding association with familiar civil society organizations, lacking leaders, and displaying little conventional organization, *los indignados*, similar to PPF, achieved high levels of WUNC.

Two broad organizational patterns characterize these increasingly common digitally enabled action networks. Some cases, such as PPF, are coordinated behind the scenes by networks of established issue advocacy organizations that step back from branding the actions in terms of particular organizations, memberships, or conventional collective action frames. Instead, they cast a broader public engagement net using interactive digital media and easy-to-personalize action themes, often deploying batteries of social technologies to help citizens spread the word over their personal networks. The second pattern, typified by the *indignados*, and the *occupy* protests in the United States, entails technology platforms and applications taking the role of established political organizations. In this network mode, political demands and grievances are often shared in very personalized accounts that travel over social networking platforms, email lists, and online coordinating platforms. For example, the easily personalized action frame 'we are the 99 per cent' that emerged from the US *occupy* protests in 2011 quickly traveled the world via personal stories and images shared on social networks such as Tumblr, Twitter, and Facebook.

Compared to many conventional social movement protests with identifiable membership organizations leading the way under common banners and collective identity frames, these more personalized, digitally mediated collective action formations have frequently been larger; have scaled up more quickly; and have been flexible in tracking moving political targets and bridging different issues. Whether we look at PPF, Arab Spring, the *indignados*, or *occupy*, we note surprising success in communicating simple political messages directly to outside publics using common digital technologies such as Facebook or Twitter. Those media feeds are often picked up as news sources by conventional journalism organizations.[5] In addition, these digitally mediated action networks often seem to be accorded higher levels of WUNC than their more conventional social movement counterparts. This observation is based on comparisons of more conventional anti-capitalist collective actions organized by movement groups, in contrast with both the organizationally enabled PPF protests and relatively more self-organizing 15M mobilizations in Spain and the Occupy Wall Street protests, which quickly spread to thousands of other places. The differences between both types of digitally mediated action and more conventional organization-centered and brokered collective actions led us to see interesting differences in underlying organizational logics and in the role of communication as an organizing principle.

The rise of digitally networked action (DNA) has been met with some understandable skepticism about what really is so very new about it, mixed with concerns about what it means for the political capacities of organized dissent. We are interested in understanding how these more personalized varieties of collective action work: how they are organized, what sustains them, and when they are politically effective. We submit that convincingly addressing such questions requires recognizing the differing logics of action that underpin distinct kinds of collective action networks. This article thus develops a conceptual framework of such logics, on the basis of which further questions about DNA may then be tackled.

We propose that more fully understanding contemporary large-scale networks of contentious action involves distinguishing between at least two logics of action that may be in play: the familiar logic of collective action and the less familiar logic of *connective* action. Doing so in turn allows us to discern three ideal action types, of which one is characterized by the familiar logic of collective action, and other two types involve more personalized action formations that differ in terms of whether formal organizations are more or less central in enabling a connective communication logic. A first step in understanding DNA, the DNA at the core of connective action, lies in defining personalized communication and its role along with digital media in the organization of what we call connective action.

Personal action frames and social media networks

Structural fragmentation and individualization in many contemporary societies constitute an important backdrop to the present discussion. Various breakdowns in group memberships and institutional loyalties have trended in the more economically developed industrial democracies, resulting from pressures of economic globalization spanning a period from roughly the 1970s through the end of the last century (Bennett 1998; Putnam 2000). These sweeping changes have produced a shift in social and political orientations among younger generations in the nations that we now term the post-industrial democracies (Inglehart 1997). These individualized orientations result in engagement with politics as an expression of personal hopes, lifestyles, and grievances. When enabled by various kinds of communication technologies, the resulting DNAs in post-industrial democracies bear some remarkable similarities to action formations in decidedly undemocratic regimes such as those swept by the Arab Spring. In both contexts, large numbers of similarly disaffected individuals seized upon opportunities to organize collectively through access to various technologies (Howard & Hussain 2011). Those connectivities fed in and out of the often intense face-to-face interactions going on in squares, encampments, mosques, and general assembly meetings.

In personalized action formations, the nominal issues may resemble older movement or party concerns in terms of topics (environment, rights, women's equality, and trade fairness) but the ideas and mechanisms for organizing action become more personalized than in cases where action is organized on the basis of social group identity, membership, or ideology. These multi-faceted processes of individualization are articulated differently in different societies, but include the propensity to develop flexible political identifications based on personal lifestyles (Giddens 1991; Inglehart 1997; Bennett 1998; Bauman 2000; Beck & Beck-Gernsheim 2002), with implications in collective action (McDonald 2002; Micheletti 2003; della Porta 2005) and organizational participation (Putnam 2000; Bimber *et al.*, in press). People may still join actions in large numbers, but the identity reference is more derived through inclusive and diverse large-scale personal expression rather than through common group or ideological identification.

This shift from group-based to individualized societies is accompanied by the emergence of flexible social 'weak tie' networks (Granovetter 1973) that enable identity expression and the navigation of complex and changing social and political landscapes. Networks have always been part of society to help people navigate life within groups or between groups, but the late modern society involves networks that become more central organizational forms that transcend groups and constitute core organizations in their own right (Castells 2000). These networks are established and scaled through various sorts of digital technologies that are by no means value neutral in enabling quite different kinds of communities to form and diverse actions to be organized, from auctions on eBay to protests in different cultural and social settings. Thus, the two elements of 'personalized communication' that we identify as particularly important in large-scale connective action formations are:

(1) Political content in the form of easily personalized ideas such as PPF in the London 2009 protests, or 'we are the 99 per cent' in the later *occupy* protests. These frames require little in the way of persuasion, reason, or reframing to bridge differences with how others may feel about a common problem. *These personal action frames are inclusive of different personal reasons for contesting a situation that needs to be changed.*

(2) Various personal communication technologies that enable sharing these themes. Whether through texts, tweets, social network sharing, or posting YouTube mashups, the communication process itself often involves further personalization through the spreading of digital connections

among friends or trusted others. Some more sophisticated custom coordinating platforms can resemble organizations that exist more online than off.

As we followed various world protests, we noticed a dazzling array of personal action frames that spread through social media. Both the acts of sharing these personal calls to action and the social technologies through which they spread help explain both how events are communicated to external audiences and how the action itself is organized. Indeed, in the limiting case, the communication network becomes the organizational form of the political action (Earl & Kimport 2011). We explore the range of differently organized forms of contention using personalized communication up to the point at which they enter the part of the range conventionally understood as social movements. This is the boundary zone in which what we refer to as connective action gives way to collective action.

The case of PPF occupies an interesting part of this range of contentious action because there were many conventional organizations involved in the mobilization, from churches to social justice NGOs. Yet, visitors to the sophisticated, stand alone, PPF coordinating platform (which served as an interesting kind of organization in itself) were not asked to pledge allegiance to specific political demands on the organizational agendas of the protest sponsors. Instead, visitors to the organizing site were met with an impressive array of social technologies, enabling them to communicate in their own terms with each other and with various political targets. The centerpiece of the PPF site was a prominent text box under an image of a megaphone that invited the visitor to 'Send Your Own Message to the G20'. Many of the messages to the G20 echoed the easy-to-personalize action frame of PPF, and they also revealed a broad a range of personal thoughts about the crisis and possible solutions.

'PPF' as a personal action frame was easy to shape and share with friends near and far. It became a powerful example of what students of viral communication refer to as a meme: a symbolic packet that travels easily across large and diverse populations because it is easy to imitate, adapt personally, and share broadly with others. Memes are network building and bridging units of social information transmission similar to genes in the biological sphere (Dawkins 1989). They travel through personal appropriation, and then by imitation and personalized expression via social sharing in ways that help others appropriate, imitate, and share in turn (Shifman, forthcoming). The simple PPF protest meme traveled interpersonally, echoing through newspapers, blogs, Facebook friend networks, Twitter streams, Flickr pages, and other sites on the Internet, leaving traces for years after the events.[6] Indeed, part of the meme traveled to Toronto more than a year later where the leading civil society groups gave the name 'People First' to their demonstrations. And many people in the large crowds in Seoul in the last G20 meeting of the series could be seen holding up red and white 'PPF' signs in both English and Korean (Weller 2010).

Something similar happened in the case of the *indignados*, where protesters raised banners and chants of 'Shhh . . . the Greeks are sleeping', with reference to the crushing debt crisis and severe austerity measures facing that country. This idea swiftly traveled to Greece where Facebook networks agreed to set alarm clocks at the same time to wake up and demonstrate. Banners in Athens proclaimed: 'We've awakened! What time is it? Time for them to leave!' and 'Shhh . . . the Italians are sleeping' and 'Shhh . . . the French are sleeping'. These efforts to send personalized protest themes across national and cultural boundaries met with varying success, making for an important cautionary point: We want to stress that not all personal action frames travel equally well or equally far. The fact that these messages traveled more easily in Spain and Greece than in France or Italy is an interesting example pointing to the need to study failures as well as successes. Just being easy to personalize (e.g. I am personally indignant about x, y, and z, and so I join with *los indignados*) does not ensure successful diffusion. Both political opportunities and conditions for social adoption may differ from situation to situation. For example, the limits in the Italian case may reflect an already established popular antigovernment network centered on comedian/activist Bepe Grillo. The French case may involve the ironic efforts of established groups on the left to lead incipient solidarity protests with the *indignados*, and becoming too heavy handed in suggesting messages and action programs.

Personal action frames do not spread automatically. People must show each other how they can appropriate, shape, and share themes. In this interactive process of personalization and sharing, communication networks may become scaled up and stabilized through the digital technologies people use to share ideas and relationships with others. These technologies and their use patterns often remain in place as organizational mechanisms. In the PPF and the *indignados* protests, the communication processes themselves represented important forms of organization.

In contrast to personal action frames, other calls to action more clearly require joining with established groups or ideologies. These more conventionally understood collective action frames are more likely to stop at the edges of communities, and may require resources beyond communication technologies to bridge the gaps or align different collective frames (Snow & Benford 1988; Benford & Snow 2000). For example, another set of protests in London at the start of the financial crisis was organized by a coalition of more radical groups under the name G20 Meltdown. Instead of mobilizing the expression of large-scale personal concerns, they demanded ending the so-called neoliberal economic policies of the G20, and some even called for the end to capitalism itself. Such demands typically come packaged with more demanding calls to join in particular repertoires of a collective action. Whether those repertoires are violent or non-violent, they typically require adoption of shared ideas and behaviors. These anarcho-socialist demonstrations drew on familiar anti-capitalist slogans and calls to 'storm the banks' or 'eat the rich' while staging dramatic marches behind the four horsemen of the economic apocalypse riding from the gates of old London to the Bank of England. These more radical London events drew smaller turnouts (some 5,000 for the Bank of England march and 2,000 for a climate encampment), higher levels of violence, and generally negative press coverage (Bennett & Segerberg 2011). While scoring high on commitment in terms of the personal costs of civil disobedience and displaying unity around anti-capitalist collective action frames, these demonstrations lacked the attributions of public worthiness (e.g. recognition from public officials, getting their messages into the news) and the numbers that gave PPF its higher levels of WUNC.

Collective action frames that place greater demands on individuals to share common identifications or political claims can also be regarded as memes, in the sense that slogans such as 'eat the rich' have rich histories of social transmission. This particular iconic phrase may possibly date to Rousseau's quip: 'When the people shall have nothing more to eat, they will eat the rich.' The crazy course of that meme's passage down through the ages includes its appearance on T-shirts in the 1960s and in rock songs of that title by Aerosmith and Motorhead, just to scratch the surface of its history of travel through time and space, reflecting the sequence of appropriation, personal expression, and sharing. One distinction between personal action and collective action memes seems to be that the latter require somewhat more elaborate packaging and ritualized action to reintroduce them into new contexts. For example, the organizers of the 'Storm the Banks' events staged an elaborate theatrical ritual with carnivalesque opportunities for creative expression as costumed demonstrators marched behind the Four Horsemen of the financial apocalypse.[7] At the same time, the G20 Meltdown discourse was rather closed, requiring adopters to make common cause with others. The Meltdown coalition had an online presence, but they did not offer easy means for participants to express themselves in their own voice (Bennett & Segerberg 2011). This suggests that more demanding and exclusive collective action frames can also travel as memes, but more often they hit barriers at the intersections of social networks defined by established political organizations, ideologies, interests, class, gender, race, or ethnicity. These barriers often require resources beyond social technologies to overcome.

While the idea of memes may help to focus differences in transmission mechanisms involved in more personal versus collective framing of action, we will use the terms personal action frames and collective action frames as our general concepts. This conceptual pairing locates our work alongside analytical categories used by social movement scholars (Snow & Benford 1988; Benford & Snow 2000). As should be obvious, the differences we are sketching between personal and collective action frames are not about being online versus offline. All contentious action networks are in important ways embodied and enacted by people on the ground (Juris 2008; Routledge & Cumbers 2009). Moreover, most formal political organizations have discovered that the growing sophistication and ubiquity of social media can reduce the resource costs of public outreach and coordination, but these uses of media do not change the action dynamics by altering the fundamental principles of organizing collectivities. By contrast, digital media networking can change the organizational game, given the right interplay of technology, personal action frames, and, when organizations get in the game, their willingness to relax collective identification requirements in favor of personalized social networking among followers.

The logic of collective action that typifies the modern social order of hierarchical institutions and membership groups stresses the organizational dilemma of getting individuals to overcome resistance to joining actions where personal participation costs may outweigh marginal gains, particularly when people can ride on the efforts of others for free, and reap the benefits if those others win the day. In short, conventional collective action typically requires people to make more difficult choices and adopt more self-changing social identities than DNA based on personal action frames organized around social technologies. The spread of collective identifications typically requires more education, pressure, or socialization, which in turn makes higher demands on formal organization and resources such as money to pay rent for organization offices, to generate publicity, and to hire professional staff organizers (McAdam *et al.* 1996).[8] Digital media may help reduce some costs in these processes, but they do not fundamentally change the action dynamics.

As noted above, the emerging alternative model that we call *the logic of connective action* applies increasingly to life in late modern societies in which formal organizations are losing their grip on individuals, and group ties are being replaced by large-scale, fluid social networks (Castells 2000).[9] These networks can operate importantly through the organizational processes of social media, and their logic does not require strong organizational control or the symbolic construction of a united 'we'. The logic of connective action, we suggest, entails a dynamic of its own and thus deserves analysis on its own analytical terms.

Two logics: collective and connective action

Social movements and contentious politics extend over many different kinds of phenomena and action (Melucci 1996; McAdam *et al.* 2001; Tarrow 2011). The talk about new forms of collective action may reflect ecologies of action that are increasingly complex (Chesters & Welsh 2006). Multiple organizational forms operating within such ecologies may be hard to categorize, not least because they may morph over time or context, displaying hybridity of various kinds (Chadwick 2011). In addition, protest and organizational work is occurring both online and off, using technologies of different capabilities, sometimes making the online/offline distinction relevant, but more often not (Earl & Kimport 2011; Bimber *et al.*, in press).

Some mark a turning point in patterns of contemporary contentious politics, which mix different styles of organization and communication, along with the intersection of different issues with the iconic union of 'teamsters and turtles' in the Battle of Seattle in 1999, during which burly union members marched alongside environmental activists wearing turtle costumes in battling a rising neoliberal trade regime that was seen as a threat to democratic control of both national economies and the world environment. Studies of such events show that there are still plenty of old-fashioned meetings, issue brokering, and coalition building going on (Polletta 2002). At the same time, however, there is increasing coordination of action by organizations and individuals using digital media to create networks, structure activities, and communicate their views directly to the world. This means that there is also an important degree of technology-enabled networking (Livingston & Asmolov 2010) that makes highly personalized, socially mediated communication processes fundamental structuring elements in the organization of many forms of connective action.

How do we sort out what organizational processes contribute what qualities to collective and connective action networks? How do we identify the borders between fundamentally different types of action formations: that is, what are the differences between collective and connective action, and where are the hybrid overlaps? We propose a starting point for sorting out some of the complexity and overlap in the forms of action by distinguishing between two logics of action. The two logics are associated with distinct dynamics, and thus draw attention to different dimensions for analysis. It is important to separate them analytically as the one is less familiar than the other, and this in turn constitutes an important stumbling block for the study of much contemporary political action that we term *connective action*.[10]

The more familiar action logic is the *logic of collective action,* which emphasizes the problems of getting individuals to contribute to the collective endeavor that typically involves seeking some sort of public good (e.g. democratic reforms) that may be better attained through forging a common cause. The classical formulation of this problem was articulated by Olson (1965), but the implications of his general logic have reached far beyond the original formulation. Olson's intriguing observation was that people in fact cannot be expected to act together just because they share a common problem or goal. He held that in large groups in which individual contributions are less noticeable, rational individuals will free-ride on the efforts of others: it is more cost-efficient not to contribute if you can enjoy the good without contributing. Moreover, if not enough people join in creating the good your efforts are wasted anyway. Either way, it is individually rational not to contribute, even if all agree that all would be better off if everyone did. This thinking fixes attention on the problematic dynamics attending the rational action of atomistic individuals, and at the same time makes resource-rich organizations a central concern. Both the solutions Olson discerned – coercion and selective incentives – implied organizations with substantial capacity to monitor, administer, and distribute such measures.

In this view, formal organizations with resources are essential to harnessing and coordinating individuals in common action. The early application of this logic to contentious collective action was most straightforwardly exemplified in resource mobilization theory (RMT), in which social movement scholars explicitly adopted Olson's framing of the collective action problem and its organization-centered solution. Part of a broader wave rejecting the idea of social movements as irrational behavior erupting out of social dysfunction, early RMT scholars accepted the problem of rational free-riders as a fundamental challenge and regarded organizations and their ability to mobilize resources as critical elements of social movement success. Classic formulations came from McCarthy and Zald (1973, 1977) who theorized the rise of external support and resources available to social movement organizations (SMOs), and focused attention on the professionalization of movement organizations and leaders in enabling more resource-intensive mobilization efforts.

The contemporary social movement field has moved well beyond the rational choice orientation of such earlier work. Indeed, important traditions developed independently of, or by rejecting, all or parts of the resource mobilization perspective and by proposing that we pay more attention to the role of identity, culture, emotion, social networks, political process, and opportunity structures (Melucci 1996; McAdam *et al.* 2001; della Porta & Diani 2006). We do not suggest that these later approaches cling to rational choice principles. We do, however, suggest that echoes of the modernist logic of collective action can still be found to play a background role even in work that is in other ways far removed from the rational choice orientation of Olson's original argument. This comes out in assumptions about the importance of particular forms of organizational coordination and identity in the attention given to organizations, resources, leaders, coalitions, brokering differences, cultural or epistemic communities, the importance of formulating collective action frames, and bridging of differences among those frames. Connective action networks may vary in terms of stability, scale, and coherence, but they are organized by different principles. Connective action networks are typically far more individualized and technologically organized sets of processes that result in action without the requirement of collective identity framing or the levels of organizational resources required to respond effectively to opportunities.

One of the most widely adopted approaches that moved social movement research away from the rational choice roots toward a more expansive collective action logic is the analysis of collective action frames, which centers on the processes of negotiating common interpretations of collective identity linked to the contentious issues at hand (Snow *et al.* 1986; Snow & Benford 1988; Hunt *et al.* 1994; Benford & Snow 2000). Such framing work may help to mobilize individuals and ultimately lower resource costs by retaining their emotional commitment to action. At the same time, the formulation of ideologically demanding, socially exclusive, or high conflict collective frames also invites fractures, leading to an analytical focus on how organizations manage or fail to bridge these differences. Resolving these frame conflicts may require the mobilization of resources to bridge differences between groups that have different goals and ways of understanding their issues. Thus, while the evolution of different strands of social movement theory has moved away from economic collective action models, many still tend to emphasize the importance of organizations that have strong ties to members and followers, and the resulting ways in which collective identities are forged and fractured among coalitions of those organizations and their networks.

Sustainable and effective collective action from the perspective of the broader logic of collective action typically requires varying levels of organizational resource mobilization deployed in organizing, leadership, developing common action frames, and brokerage to bridge organizational differences. The opening or closing of political opportunities affects this resource calculus (Tarrow 2011), but overall, large-scale action networks that reflect this collective action logic tend to be characterized in terms of numbers of distinct groups networking to bring members and affiliated participants into the action and to keep them there. On the individual level, collective action logic emphasizes the role of social network relationships and connections as informal preconditions for more centralized mobilization (e.g. in forming and spreading action frames, and forging common identifications and relations of solidarity and trust). At the organizational level, the strategic work of brokering and bridging coalitions between organizations with different standpoints and constituencies becomes the central activity for analysis (cf. Diani 2011). Since the dynamics of action in networks characterized by this logic tends not to change significantly with digital media, it primarily invites analysis of how such tools help actors do what they were already doing (cf. Bimber *et al.* 2009; Earl & Kimport 2011).

Movements and action networks characterized by these variations on the logic of collective action are clearly visible in contemporary society. They have been joined by many other mobilizations that may superficially seem like movements, but on closer inspection lack many of the traditional defining characteristics. Efforts to push these kinds of organization into recognizable social movement categories diminish our capacity to understand one of the most interesting developments of our times: how fragmented, individualized populations that are hard to reach and even harder to induce to share personally transforming collective identities somehow find ways to mobilize protest networks from Wall Street to Madrid to Cairo. Indeed, when people are individualized in their social orientations, and thus structurally or psychologically unavailable to modernist forms of political movement organization, resource mobilization becomes increasingly costly and has diminishing returns. Organizing such populations to overcome free riding and helping them to shape identities in common is not necessarily the most successful or effective logic for organizing collective action. When people who seek more personalized paths to concerted action are familiar with practices of social networking in everyday life, and when they have access to technologies from mobile phones to computers, they are already familiar with a different logic of organization: the logic of connective action.

The *logic of connective action* foregrounds a different set of dynamics from the ones just outlined. At the core of this logic is the recognition of digital media as organizing agents. Several collective action scholars have explored how digital communication technology alters the parameters of Olson's original theory of collective action. Lupia and Sin (2003) show how Olson's core assumption about weak individual commitment in large groups (free riding) may play out differently under conditions of radically reduced communication costs. Bimber *et al*. (2005) in turn argue that public goods themselves may take on new theoretical definition as erstwhile free-riders find it easier to become participants in political networks that diminish the boundaries between public and private – boundaries that are blurred in part by the simultaneous public/private boundary crossing of ubiquitous social media.

Important for our purposes here is the underlying economic logic of digitally mediated social networks as explained most fully by Benkler (2006). He proposes that participation becomes self-motivating as personally expressive content is shared with, and recognized by, others who, in turn, repeat these networked sharing activities. When these interpersonal networks are enabled by technology platforms of various designs that coordinate and scale the networks, the resulting actions can resemble collective action, yet without the same role played by formal organizations or transforming social identifications. In place of content that is distributed and relationships that are brokered by hierarchical organizations, social networking involves co-production and co-distribution, revealing a different economic and psychological logic: co-production and sharing based on personalized expression. This does not mean that all online communication works this way. Looking at most online newspapers, blogs, or political campaign sites makes it clear that the logic of the organization-centered brick and mortar world is often reproduced online, with little change in organizational logic beyond possible efficiency gains (Bimber & Davis 2003; Foot & Schneider 2006). Yet, many socially mediated networks do operate with an alternative logic that also helps to explain why people labor collectively for free to create such things as open source software, Wikipedia, WikiLeaks, and the Free and Open Source Software that powers many protest networks (Calderaro 2011).

In this connective logic, taking public action or contributing to a common good becomes an act of personal expression and recognition or self-validation achieved by sharing ideas and actions in trusted relationships. Sometimes the people in these exchanges may be on the other side of the world, but they do not require a club, a party, or a shared ideological frame to make the connection. In place of the initial collective action problem of getting the individual to contribute, the starting point of connective action is the self-motivated (though not necessarily self-centered) sharing of already internalized or personalized ideas, plans, images, and resources with networks of others. This 'sharing' may take place in networking sites such as Facebook, or via more public media such as Twitter and YouTube through, for example, comments and re-tweets.[11] Action networks characterized by this logic may scale up rapidly through the combination of easily spreadable personal action frames and digital technology enabling such communication. This invites analytical attention to the network as an organizational structure in itself.

Technology-enabled networks of personalized communication involve more than just exchanging information or messages. The flexible, recombinant nature of DNA makes these web spheres and their offline extensions more than just communication systems. Such networks are flexible organizations in themselves, often enabling coordinated adjustments and rapid action aimed at often shifting political targets, even crossing geographic and temporal boundaries in the process. As Diani (2011) argues, networks are not just precursors or building blocks of collective action: they are in themselves organizational structures that can transcend the elemental units of organizations and individuals.[12] As noted earlier, communication technologies do not change the action dynamics in large-scale networks characterized by the logic of collective action. In the networks characterized by connective action, they do.

The organizational structure of people and social technology emerges more clearly if we draw on the actor-network theory of Latour (2005) in recognizing digital networking mechanisms (e.g. various social media and devices that run them) as potential network agents alongside human actors (i.e. individuals and organizations). Such digital mechanisms may include: organizational connectors (e.g. web links), event coordination (e.g. protest calendars), information sharing (e.g. YouTube and Facebook), and multifunction networking platforms in which other networks become embedded (e.g. links in Twitter and Facebook posts), along with various capacities of the devices that run them. These technologies not only create online meeting places and coordinate offline activities, but they also help calibrate relationships by establishing levels of transparency, privacy, security, and interpersonal trust. It is also important that these digital traces may remain behind on the web to provide memory records or action repertoires that might be passed on via different mechanisms associated with more conventional collective action such as rituals or formal documentation.

The simple point here is that collective and connective logics are distinct logics of action (both in terms of identity and choice processes), and thus both deserve analysis on their own terms. Just as traditional collective action efforts can fail to result in sustained or effective movements, there is nothing preordained about the results of digitally mediated networking processes. More often than not, they fail badly. The transmission of personal expression across networks may or may not become scaled up, stable, or capable of various kinds of targeted action depending on the kinds of social technology designed and appropriated by participants, and the kinds of opportunities that may motivate anger or compassion across large numbers of individuals. Thus, the Occupy Wall Street protests that spread in a month from New York to over 80 countries and 900 cities around the world might not have succeeded without the inspiring models of the Arab Spring or the *indignados* in Spain, or the worsening economic conditions that provoked anger among increasing numbers of displaced individuals. Yet, when the 'Occupy' networks spread under the easy-to-personalize action frame of 'we are the 99 per cent', there were few identifiable established political organizations at the center of them. There was even a conscious effort to avoid designating leaders and official spokespeople. The most obvious organizational forms were the layers of social technologies and websites that carried news reported by participants and displayed tools for personalized networking. One of the sites was '15.10.11 united for #global change'.[13] Instead of the usual 'who are we' section of the website, #globalchange asked: 'who are you?'.

Collective and connective action may co-occur in various formations within the same ecology of action. It is nonetheless possible to discern three clear ideal types of large-scale action networks. While one is primarily characterized by collective action logic, the other two are connective action networks distinguished by the role of formal organizations in facilitating personalized engagement. As noted above, conventional organizations play a less central role than social technologies in relatively self-organizing networks such as the *indignados* of Spain, the Arab Spring uprisings, or the '*occupy*' protests that spread from Wall Street around the world. In contrast to these more technology-enabled networks, we have also observed hybrid networks (such as PPF) where conventional organizations operate in the background of protest and issue advocacy networks to enable personalized engagement. This hybrid form of organizationally enabled connective action sits along a continuum somewhere between the two ideal types of conventional organizationally managed collective action and relatively more self-organized connective action. The following section presents the details of this three-part typology. It also suggests that co-existence, layering, and movement across the types become an important part of the story.

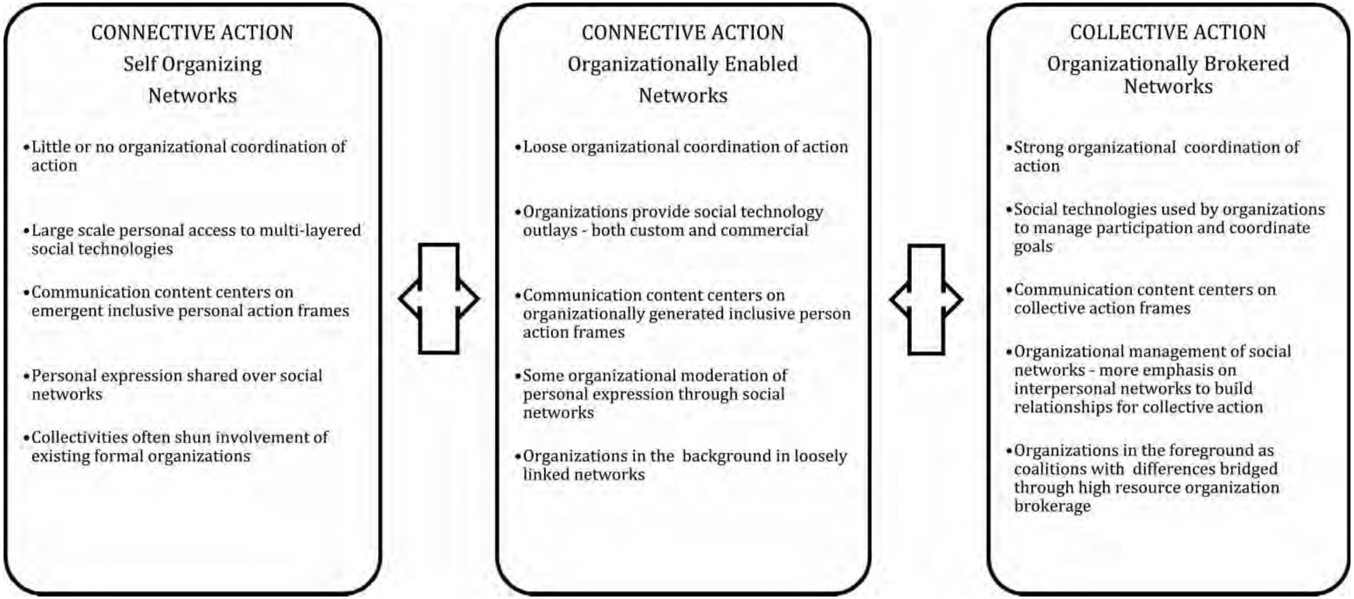
A typology of collective and connective action networks

We draw upon these distinct logics of action (and the hybrid form that reveals a tension between them) to develop a three-part typology of large-scale action networks that feature prominently in contemporary contentious politics. One type represents the brokered organizational networks characterized by the logic of collective action, while the others represent two significant variations on networks primarily characterized by the logic of connective action. All three models may explain differences between and dynamics within large-scale action networks in event-centered contention, such as protests and sequences of protests as in the examples we have already discussed. They may also apply to more stable issue advocacy networks that engage people in everyday life practices supporting causes outside of protest events such as campaigns. The typology is intended as a broad generalization to help understand different dynamics. None of the types are exhaustive social movement models. Thus, this is not an attempt to capture, much less resolve, the many differences among those who study social movements. We simply want to highlight the rise of two forms of digitally networked connective action that differ from some common assumptions about collective action in social movements, and, in particular, that rely on mediated networks for substantial aspects of their organization.

Figure 1 presents an overview of the two connective action network types and contrasts their organizational properties with more familiar collective action network organizational characteristics. The ideal collective action type at the right side in the figure describes large-scale action networks that depend on brokering organizations to carry the burden of facilitating cooperation and bridging differences when possible. As the anti-capitalist direct action groups in the G20 London Summit protests exemplified, such organizations will tend to promote more exclusive collective action frames that require frame bridging if they are to grow. They may use digital media and social technologies more as means of mobilizing and managing participation and coordinating goals, rather than inviting personalized interpretations of problems and self-organization of action. In addition to a number of classic social movement accounts (e.g. McAdam 1986), several of the NGO networks discussed by Keck and Sikkink (1998) also accord with this category (Bennett 2005).

At the other extreme on the left side in the figure we place connective action networks that self-organize largely without central or ‘lead’ organizational actors, using technologies as important organizational agents. While some formal organizations may be present, they tend to remain at the periphery or may exist as much in online as in offline forms. In place of collective action frames, personal action frames become the transmission units across trusted social networks. The loose coordination of the *indignados* exemplifies this ideal type, with conventional organizations deliberately kept at the periphery as easily adapted personal action frames travel online and offline with the aid of technology platforms such as the *Democracia Real Ya!* organization.[14]

In between the organizationally brokered collective action networks and the more self-organizing (technology organized) connective action network is the hybrid pattern introduced above. This middle type involves formal organizational



CONNECTIVE ACTION
Self Organizing
Networks

- Little or no organizational coordination of action
- Large scale personal access to multi-layered social technologies
- Communication content centers on emergent inclusive personal action frames
- Personal expression shared over social networks
- Collectivities often shun involvement of existing formal organizations

CONNECTIVE ACTION
Organizationally Enabled
Networks

- Loose organizational coordination of action
- Organizations provide social technology outlays - both custom and commercial
- Communication content centers on organizationally generated inclusive person action frames
- Some organizational moderation of personal expression through social networks
- Organizations in the background in loosely linked networks

COLLECTIVE ACTION
Organizationally Brokered
Networks

- Strong organizational coordination of action
- Social technologies used by organizations to manage participation and coordinate goals
- Communication content centers on collective action frames
- Organizational management of social networks - more emphasis on interpersonal networks to build relationships for collective action
- Organizations in the foreground as coalitions with differences bridged through high resource organization brokerage

FIGURE 1 Elements of connective and collective action networks.

actors stepping back from projecting strong agendas, political brands, and collective identities in favor of using resources to deploy social technologies enabling loose public networks to form around personalized action themes. The middle type may also encompass more informal organizational actors that develop some capacities of conventional organizations in terms of resource mobilization and coalition building without imposing strong brands and collective identities.[15] For example, many of the General Assemblies in the *occupy* protests became resource centers, with regular attendance, division of labor, allocation of money and food, and coordination of actions. At the same time, the larger communication networks that swirled around these protest nodes greatly expanded the impact of the network. The surrounding technology networks invited loose tied participation that was often in tension with the face-to-face ethos of the assemblies, where more committed protesters spent long hours with dwindling numbers of peers debating on how to expand participation without diluting the levels of commitment and action that they deemed key to their value scheme. Thus, even as *occupy* displayed some organizational development, it was defined by its self-organizing roots.

Networks in this hybrid model engage individuals in causes that might not be of such interest if stronger demands for membership or subscribing to collective demands accompanied the organizational offerings. Organizations facilitating these action networks typically deploy an array of custom built (e.g. 'send your message') and outsourced (e.g. Twitter) communication technologies. This pattern fit the PPF demonstrations discussed earlier, where some 160 civil society organizations – including major NGOs such as Oxfam, Tearfund, Catholic Relief, and World Wildlife Fund – stepped back from their organizational brands to form a loose social network inviting publics to engage with each other and take action. They did this even as they negotiated with other organizations over such things as separate days for the protests (Bennett & Segerberg 2011).

The formations in the middle type reflect the pressures that Bimber *et al.* (2005) observed in interest organizations that are suffering declining memberships and have had to develop looser, more entrepreneurial relations with followers. Beyond the ways in which particular organizations use social technologies to develop loose ties with followers, many organizations also develop loose ties with other organizations to form vast online networks sharing and bridging various causes. Although the scale and complexity of these networks differ from the focus of Granovetter's (1973) observations about the strength of weak ties in social networks, we associate this idea with the elements of connective action: the loose organizational linkages, technology deployments, and personal action frames. In observing the hybrid pattern of issue advocacy organizations facilitating personalized protest networks, we traced a number of economic justice and environmental networks, charting protests, campaigns, and issue networks in the UK, Germany, and Sweden (Bennett & Segerberg, forthcoming).[16] In each case, we found (with theoretically interesting variations) campaigns, protest events, and everyday issue advocacy networks that displayed similar organizational signatures: (a) familiar NGOs and other civil society organizations joining loosely together to provide something of a networking backbone, (b) for digital media networks engaging publics with contested political issues, yet with (c) remarkably few efforts to brand the issues around specific organizations, own the messages, or control the understandings of individual participants. The organizations had their political agendas on offer, to be sure, but, as members of issue networks, put the public face on the individual citizen and provided social technologies to enable personal engagement through easy-to-share images and personal action frames.

The organizations that refrain from strongly branding their causes or policy agendas in this hybrid model do not necessarily give up their missions or agendas as name brand public advocacy organizations. Instead, some organizations interested in mobilizing large and potentially WUNC-y publics in an age of social networking are learning to shift among different organizational repertoires, morphing from being hierarchical, mission-driven NGOs in some settings to being facilitators in loosely linked public engagement networks in others. As noted by Chadwick (2007, 2011), organizational hybridity makes it difficult to apply fixed categories to many organizations as they variously shift from being issue advocacy NGOs to policy think tanks, to SMOs running campaigns or protests, to multi-issue organizations, to being networking hubs for connective action. In other words, depending on when, where, and how one observes an organization, it may appear differently as an NGO, SMO, INGO, TNGO, NGDO (non governmental organization, social movement organization, international non governmental organization, transnational non governmental organization, non governmental development organization), an interest advocacy group, a political networking hub, and so on. Indeed, one of the advantages of seeing the different logics at play in our typology is to move away from fixed categorization schemes, and observe actually occurring combinations of different types of action within complex protest ecologies, and shifts in dominant types in response to events and opportunities over time.

The real world is of course far messier than this three-type model. In some cases, we see action formations corresponding to our three models side by side in the same action space. The G20 London protest offered a rare case in which organizationally enabled and more conventional collective action were neatly separated over different days. More often, the different forms layer and overlap, perhaps with violence disrupting otherwise peaceful mobilizations as occurred in the Occupy Rome protests on 15 October 2011, and in a number of *occupy* clashes with police in the United States. In still other action cycles, we see a movement from one model to another over time. In some relatively distributed networks, we observe a pattern of informal *organizational resource seeking*, in which informal organizational resources and communication spaces are linked and shared (e.g. re-tweeted), enabling emergent political concerns and goals to be nurtured without being co-opted by existing organizations and their already fixed political agendas. This pattern occurred in the self-organizing Twitter network that emerged around the 15th UN climate summit in Copenhagen. As the long tail of that network handed its participants off to the Twitter stream devoted to the next summit in Cancun, we saw an increase in links to organizations of various kinds, along with growing links to and among climate bloggers (Segerberg & Bennett 2011). Such variations on different organizational forms offer intriguing opportunities for further analyses aimed at explaining whether mobilizations achieve various goals, and attain different levels of WUNC.

In these varying ways, personalized connective action networks cross paths (sometimes with individual organizations morphing in the process) with more conventional collective action networks centered on SMOs, interest organizations, and brand-conscious NGOs. As a result, while we argue that these networks are an organizational form in themselves, they are often hard to grasp and harder to analyze because they do not behave like formal organizations. Most formal organizations are centered (e.g. located in physical space), hierarchical, bounded by mission and territory, and defined by relatively known and countable memberships (or in the case of political parties, known and reachable demographics). By contrast, many of today's issue and cause networks are relatively de-centered (constituted by multiple organizations and many direct and cyber activists), distributed, or flattened organizationally as a result of these multiple centers, relatively unbounded, in the sense of crossing both geographical and issue borders, and dynamic in terms of the changing populations who may opt in and out of play as different engagement opportunities are presented (Bennett 2003, 2005). Understanding how connective action engages or fails to engage diverse populations constitutes part of the analytical challenge ahead.

Compared to the vast number of theoretically grounded studies on social movement organizing, there is less theoretical work that helps explain the range of collective action formations running from relatively self-organizing to organizationally enabled connective action networks. While there are many descriptive and suggestive accounts of this kind of action, many of them insightful (e.g. Castells 2000; Rheingold 2002), we are concerned that the organizational logic and underlying dynamic of such action is not well established. It is important to gain clearer understandings of how such networks function and what organizing principles explain their growing prominence in contentious politics.

Conclusion

DNA is emerging during a historic shift in late modern democracies in which, most notably, younger citizens are moving away from parties, broad reform movements, and ideologies. Individuals are relating differently to organized politics, and many organizations are finding they must engage people differently: they are developing relationships to publics as affiliates rather than members, and offering them personal options in ways to engage and express themselves. This includes greater choice over contributing content, and introduces micro-organizational resources in terms of personal networks, content creation, and technology development skills. Collective action based on exclusive collective identifications and strongly tied networks continues to play a role in this political landscape, but this has become joined by, interspersed with, and in some cases supplanted by personalized collective action formations in which digital media become integral organizational parts. Some of the resulting DNA networks turn out to be surprisingly nimble, demonstrating intriguing flexibility across various conditions, issues, and scale.

It has been tempting for some critics to dismiss participation in such networks as noise, particularly in reaction to sweeping proclamations by enthusiasts of the democratic and participatory power of digital media. Whether from digital enthusiasts or critics, hyperbole is unhelpful. Understanding the democratic potential and effectiveness of instances of connective and collective action requires careful analysis. At the same time, there is often considerably more going on in DNA than clicktivism or facile organizational outsourcing of social networking to various commercial sites.[17] The key point of our argument is that fully explaining and understanding such action and contention requires more than just adjusting the classic social movement collective action schemes. Connective action has a logic of its own, and thus attendant dynamics of its own. It deserves analysis on its own terms.

The linchpin of connective action is the formative element of 'sharing': the personalization that leads actions and content to be distributed widely across social networks. Communication technologies enable the growth and stabilization of network structures across these networks. Together, the technological agents that enable the constitutive role of sharing in these contexts displace the centrality of the free-rider calculus and, with it, by extension, the dynamic that flows from it – most obviously, the logical centrality of the resource-rich organization. In its stead, connective action brings the action dynamics of recombinant networks into focus, a situation in which networks and communication become something more than mere preconditions and information. What we observe in these networks are applications of communication technologies that contribute an organizational principle that is different from notions of collective action based on core assumptions about the role of resources, networks, and collective identity. We call this different structuring principle the logic of connective action.

Developing ways to analyze connective action formations will give us more solid grounds for returning to the persistent questions of whether such action can be politically effective and sustained (Tilly 2004; Gladwell 2010; Morozov 2011). Even as the contours of political action may be shifting, it is imperative to develop means of thinking meaningfully about the capacities of sustainability and effectiveness in relation to connective action and to gain a systematic understanding of how such action plays out in different contexts and conditions.

The string of G20 protests surrounding the world financial crisis illustrate that different organizational strategies played out in different political settings produce a wide range of results. The protests at the Pittsburgh and Toronto G20 summits of 2009 and 2010, respectively, were far more chaotic and displayed far less WUNC than those organized under the banner of PPF in London. Disrupted by police assaults and weak organizational coordination, the Pittsburgh protests displayed a cacophony of political messages that were poorly translated in the press and even became the butt of late night comedy routines. *The Daily Show* sent a correspondent to Pittsburgh and reported on a spectrum of messages that included: a Free Tibet matching cymbal band, Palestinian peace advocates, placards condemning genocide in Darfur, hemp and marijuana awareness slogans, and denunciations of the beef industry, along with the more expected condemnations of globalization and capitalism. One protester carried a sign saying 'I protest everything', and another dressed as Batman stated that he was protesting the choice of Christian Bale to portray his movie hero. The correspondent concluded that the Pittsburgh protests lacked unity of focus and turned for advice to some people who knew how to get the job done: members of the Tea Party. The *Daily Show* panel of Tea Party experts included a woman wearing a black Smith & Wesson holster that contained a wooden crucifix with an American flag attached. When asked what the Pittsburgh protesters were doing wrong, they all agreed that there was a message problem. One said, 'I still don't know what their message is' and another affirmed, 'Stay on message and believe what you say'. The Daily Show report cut back to show a phalanx of Darth Vader-suited riot police lined up against the protesters – according to the correspondent, the 'one single understandable talking point' in Pittsburgh (*Daily Show 2009*). Humor aside, this example poses a sharp contrast to the more orderly London PPF protests that received positive press coverage of the main themes of economic and environmental justice (Bennett & Segerberg 2011).

The challenge ahead is to understand when DNA becomes chaotic and unproductive and when it attains higher levels of focus and sustained engagement over time. Our studies suggest that differing political capacities in networks depend, among other things, on whether (a) in the case of organizationally enabled DNA, the network has a stable core of organizations sharing communication linkages and deploying high volumes of personal engagement mechanisms or (b) in the case of self-organizing DNA, the digital networks are redundant and dense with pathways for individual networks to converge, enabling viral transmission of personally appealing action frames to occur.

Attention to connective action will neither explain all contentious politics nor does it replace the model of classic collective action that remains useful for analyzing social movements. But, it does shed light on an important mode of action making its mark in contentious politics today. A model focused primarily on the dynamics of classic collective action has difficulties accounting for important elements in the Arab spring, the *indignados*, the *occupy* demonstrations, or the global protests against climate change. A better understanding of connective action promises to fill some of these gaps. Such understanding is essential if we are to attain a critical perspective on some of the prominent forms of public engagement in the digital age.

Acknowledgements

The article builds on work supported by the Swedish Research Council grant Dnr. 421-2010-2303. It benefited from comments received on earlier versions, including those presented at the ECPR General Conference 2011 and the iCS/OII Symposium ‘A Decade of Internet Time’. The authors particularly wish to thank Bruce Bimber, Bob Boynton, Andrew Chadwick, Nils Gustafsson, Rasmus Kleis Nielsen, Annette Schnabel, Sidney Tarrow, and the anonymous referees for excellent comments.

Notes

[1] Simultaneous protests were held in other European cities with tens of thousands of demonstrators gathering in the streets of Berlin, Frankfurt, Vienna, Paris, and Rome.

[2] US Vice President Joe Biden asked for patience from understandably upset citizens while leaders worked on solutions, and the British Prime Minister at the time, Gordon Brown, said: ‘. . . the action we want to take (at the G20) is designed to answer the questions that the protesters have today’ (Vinocur & Barkin 2009).

[3] http://www.democraciarealya.es/

[4] Beyond the high volume of Spanish press coverage, the story of the *indignados* attracted world attention. BBC World News devoted no fewer than eight stories to this movement over the course of two months, including a feature on the march of one group across the country to Madrid, with many interviews and encounters in the words of the protesters themselves.

[5] For example, our analyses of the US *occupy* protests show that increased media attention to economic inequality in America was associated with the coverage of the *occupy* protests. While political elites were often reluctant to credit the occupiers with their newfound concern about

inequality, they nonetheless seemed to find the public opinion and media climate conducive to addressing the long-neglected issue.

6 A Google search of ‘put people first g20’ more than two years after the London events produced nearly 1.5 million hits, with most of them relevant to the events and issues of the protests well into 75 search pages deep.

7 We would note, however, that carnivalesque or theatrical expressions may entail strategically de-personalized forms of expression in which individuals take on other personae that often have historically or dramatically scripted qualities. We thank Stefania Milan for this comment.

8 We are not arguing here that all contemporary analyses of collective action rely on resource mobilization explanations (although some do). Our point is that whether resource assumptions are in the foreground or background, many collective action analyses typically rely on a set of defining assumptions centered on the importance of some degree of formal organization and some degree of strong collective identity that establishes common bonds among participants. These elements become more marginal in thinking about the organization of connective action.

9 While we focus primarily on cases in late modern, postindustrial democracies, we also attempt to develop theoretical propositions that may apply to other settings such as the Arab Spring, where authoritarian rule may also result in individualized populations that fall outside of sanctioned civil society organization, yet may have direct or indirect access to communication technologies such as mobile phones.

10 Routledge and Cumbers (2009) make a similar point in discussing horizontal and vertical models as useful heuristics for organizational logics in global justice networks (cf. Robinson & Tormey 2005; Juris 2008).

11 We are indebted to Bob Boynton for pointing out that this sharing occurs both in trusted friends networks such as Facebook and in more public exchange opportunities among strangers of the sort that occur on YouTube, Twitter, or blogs. Understanding the dynamics and interrelationships among these different media networks and their intersections is an important direction for research.

12 We have developed methods for mapping networks and inventorying the types of digital media that enable actions and information to flow through them. Showing how networks are constituted in part by technology enables us to move across levels of action that are often difficult to theorize. Network technologies enable thinking about individuals, organizations, and networks in one broad framework. This approach thus revises the starting points of classic collective action models,

which typically examine the relationships between individuals and organizations and between organizations. We expand this to include technologies that enable the formation of fluid action networks in which agency becomes shared or distributed across individual actors and organizations as networks reconfigure in response to changing issues and events (Bennett *et al*. 2011).

[13] http://www.15october.net (accessed 19 October 2011).

[14] We wish to emphasize that there is much face-to-face organizing work going on in many of these networks, and that the daily agendas and decisions are importantly shaped offline. However, the connectivity and flow of action coordination occurs importantly online.

[15] We thank the anonymous referee for highlighting this subtype.

[16] Our empirical investigations focused primarily on two types of networks that display local, national, and transnational reach: networks to promote economic justice via more equitable north south trade norms (*fair trade*) and networks for environmental and human protection from the effects of global warming (*climate change*). These networks display impressive levels of collective action and citizen engagement and they are likely to remain active into the foreseeable future. They often intersect by sharing campaigns in local, national, and transnational arenas. As such, these issue networks represent good cases for assessing the uses of digital technologies and different action frames (from personalized to collective) to engage and mobilize citizens, and to examine various related capacities and effects of those engagement efforts.

[17] Technology is not neutral. The question of the degree to which various collectivities have both appropriated and become dependent on the limitations of commercial technology platforms such as Flickr, Facebook, Twitter, or YouTube is a matter of considerable importance. For now, suffice it to note that at least some of the technologies and their networking capabilities are designed by activists for creating political networks and organizing action (Calderaro 2011).

References

Anduiza, E., Cristancho, C. & Sabucedo, J. (2011) 'The political protest of the outraged in Spain: what's new?'. Unpublished manuscript, used with permission.

Bauman, Z. (2000) *Liquid Modernity*, Polity, Cambridge.

Beck, U. & Beck-Gernsheim, E. (2002) *Individualization: Institutionalized Individualism and its Social and Political Consequences*, SAGE, London.

Benford, R. D. & Snow, D. A. (2000) 'Framing processes and social movements: an overview and an assessment', *Annual Review of Sociology*, vol. 26, pp. 611–639.

Benkler, Y. (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom*, Yale University Press, New Haven.

Bennett, W. L. (1998) 'The uncivic culture: communication, identity, and the rise of lifestyle politics', Ithiel de Sola Pool Lecture, American Political Science Association, published in *P.S.: Political Science and Politics*, vol. 31 (December), pp. 41–61.

Bennett, W. L. (2003) 'Communicating global activism: strengths and vulnerabilities of networked politics', *Information, Communication & Society*, vol. 6, no. 2, pp. 143–168.

Bennett, W. L. (2005) 'Social movements beyond borders: organization, communication, and political capacity in two eras of transnational activism', in *Transnational Protest and Global Activism*, eds D. della Porta & S. Tarrow, Rowman & Littlefield, Boulder, CO, pp. 203–226.

Bennett, W. L. & Segerberg, A. (2011) 'Digital media and the personalization of collective action: social technology and the organization of protests against the global economic crisis', *Information, Communication &Society*, vol. 14, pp. 770–799.

Bennett, W. L. & Segerberg, A. (forthcoming) *The Logic of Connective Action: Digital Media and the Personalization of Contentious Politics*, Cambridge University Press, New York.

Bennett, W. L., Lang, S. & Segerberg, A. (2011) 'Digital media and the organization of transnational advocacy: legitimacy and public engagement in national and EU issue networks', paper presented at International Studies Association Conference, Montreal, Canada, 16–19 March 2011.

Bimber, B. & Davis, R. (2003) *Campaigning Online: The Internet in U.S. Elections*, Oxford University Press, New York.

Bimber, B., Flanagin, A. & Stohl, C. (2005) 'Reconceptualizing collective action in the contemporary media environment', *Communication Theory*, vol. 15, pp. 389–413.

Bimber, B., Stohl, C. & Flanagin, A. (2009) 'Technological change and the shifting nature of political organization', in *Routledge Handbook of Internet Politics*, eds A. Chadwick & P. Howard, Routledge, London, pp. 72–85.

Bimber, B., Flanagin, A. & Stohl, C. (in press) *Collective Action in Organizations: Interaction and Engagement and Engagement in an Era of Technological Change*, Cambridge University Press, New York.

Calderaro, A. (2011) 'New political struggles in the network society: the case of free and open source software (FOSS) Movement', paper presented at ECPR General Conference, Reykjavik, 25–27 August 2011.

Castells, M. (2000) *The Network Society*, 2nd edn, Blackwell, Oxford.

Chadwick, A. (2007) 'Digital network repertoires and organizational hybridity', *Political Communication*, vol. 24, no. 3, pp. 283–301.

Chadwick, A. (2011) 'The hybrid media system', paper presented at ECPR General Conference, Reykjavik, Iceland, 25–27 August 2011.

Chesters, G. & Welsh, I. (2006) *Complexity and Social Movements: Multitudes at the End of Chaos*, Routledge, London.

Daily Show. (2009) 'Tea partiers advise G20 protesters', *Daily Show*, 1 October, [Online] Available at: http://www.thedailyshow.com/watch/thu-october-1-2009/tea-partiers-advise-g20-protesters (6 October 2010).

Dawkins, R. (1989) *The Selfish Gene*, Oxford University Press, Oxford.

Diani, M. (2011) *The Cement of Civil Society: Civic Networks in Local Settings*, Barcelona, unpublished manuscript.

Earl, J. & Kimport, K. (2011) *Digitally Enabled Social Change: Online and Offline Activism in the Age of the Internet*, MIT Press, Cambridge, MA.

Foot, K. & Schneider, S. (2006) *Web Campaigning*, MIT Press, Cambridge, MA.

Giddens, A. (1991) *Modernity and Self Identity: Self and Society in the Late Modern Age*, Stanford University Press, Stanford.

Gitlin, T. (1980) *The Whole World Is Watching: Mass Media in the Making & Unmaking of the New Left*, University of California Press, Berkeley.

Gladwell, M. (2010) 'Small change: why the revolution will not be tweeted', *The New Yorker*, 4 October.

Granovetter, M. (1973) 'The strength of weak ties', *American Journal of Sociology*, vol. 78, pp. 1360–1380.

Howard, P. & Hussain, M. (2011) 'The role of digital media', *Journal of Democracy*, vol. 22, no. 3, pp. 35–48.

Hunt, S., Benford, R. D. & Snow, D. A. (1994) 'Identity fields: framing processes and the social construction of movement identities', in *New Social Movements: From Ideology to Identity*, eds E. Laraña, H. Johnston & J. R. Gusfield, Temple University Press, Philadelphia, pp. 185–208.

Inglehart, R. (1997) *Modernization and Post-Modernization: Cultural, Economic and Political Change in 43 Societies*, Princeton University Press, Princeton.

Juris, J. (2008) *Networking Futures: The Movements against Corporate Globalization*, Duke University Press, Durham, NC.

Keck, M. & Sikkink, K. (1998) *Activists Beyond Borders: Advocacy Networks in International Politics*, Cornell University Press, Ithaca, NY.

Latour, B. (2005) *Reassembling the Social: An Introduction to Actor-Network-Theory*, Oxford University Press, Oxford.

Livingston, S. & Asmolov, G. (2010) 'Networks and the future of foreign affairs reporting', *Journalism Studies*, vol. 11, no. 5, pp. 745–760.

Lupia, A. & Sin, G. (2003) 'Which public goods are endangered? How evolving communication technologies affect "The Logic of Collective Action"', *Public Choice*, vol. 117, pp. 315–331.

McAdam, D. (1986) 'Recruitment to high-risk activism: The case of freedom summer', *American Journal of Sociology*, vol. 92, pp. 64–90.

McAdam, D., McCarthy, J. D. & Zald, M. N. (eds) (1996) 'Introduction: opportunities, mobilizing structures, and framing processes – toward a synthetic, comparative perspective on social movements', in *Comparative Perspectives on Social Movements: Political Opportunities, Mobilizing Structures, and Cultural Framings*, Cambridge University Press, New York.

McAdam, D., Tarrow, S. & Tilly, C. (2001) *Dynamics of Contention*, Cambridge University Press, New York.

McCarthy, J. D. & Zald, M. N. (1973) *The Trend of Social Movements in America: Professionalization and Resource Mobilization*, General Learning Press, Morristown, NJ.

McCarthy, J. D. & Zald, M. N. (1977) 'Resource mobilization and social movements: a partial theory', *American Journal of Sociology*, vol. 82, no. 6, pp. 1212–1241.

McDonald, K. (2002) 'From solidarity to fluidarity: social movements beyond "collective identity" – the case of globalization conflicts', *Social Movement Studies*, vol. 1, no. 2, pp. 109–128.

Melucci, A. (1996) *Challenging Codes: Collective Action in the Information Age*, Cambridge University Press, Cambridge.

Micheletti, M. (2003) *Political Virtue and Shopping*, Palgrave, New York.

Morozov, E. (2011) *The Net Delusion: How Not to Liberate the World*, Allen Lane, London.

Olson, M. (1965) *The Logic of Collective Action: Public Goods and the Theory of Groups*, Harvard University Press, Cambridge, MA.

Polletta, F. (2002) *Freedom Is an Endless meeting. Democracy in American Social Movements*, University of Chicago Press, Chicago.

della Porta, D. (2005) 'Multiple belongings, flexible identities and the construction of "another politics": between the European social forum and the local social fora', in *Transnational Protest and Global Activism*, eds D. della Porta & S. Tarrow, Rowman & Littlefield, Boulder, CO, pp. 175–202.

della Porta, D. & Diani, M. (2006) *Social Movements: An Introduction* , 2nd edn, Blackwell, Malden, MA.

Putnam, R. (2000) *Bowling Alone: The Collapse and Revival of American Community*, Simon & Schuster, New York.

Put People First (2009) [Online] Available at: http://www.putpeoplefirst.org.uk/ (6 July 2011).

Rheingold, H. (2002) *Smart Mobs: The Next Social Revolution*, Perseus Pub., Cambridge, MA.

Robinson, A. & Tormey, S. (2005) 'Horizontals, Verticals and the Conflicting Logics of Transformative Politics', in *Confronting Globalization*, eds C. el-Ojeili & P. Hayden, Palgrave, London, pp. 208–226.

Routledge, P. & Cumbers, A. (2009) *Global Justice Networks: Geographies of Transnational Solidarity*, Manchester University Press, Manchester, UK.

rtve (2011), 'Mas de seis millones de Espanoles han participado en el movimiento 15M', 6 August, [Online] Available at: http://www.rtve.es/noticias/20110806/mas-seis-millones-espanoles-han-participado-movimiento-15m/452598.shtml (18 September 2011).

Segerberg, A. & Bennett, W. L. (2011) 'Social media and the organization of collective action: using Twitter to explore the ecologies of two climate change protests', *The Communication Review*, vol. 14, no. 3, pp. 197–215.

Shifman, L. (forthcoming) *Internet Memes*, MIT Press, Cambridge, MA.

Snow, D. A. & Benford, R. D. (1988) 'Ideology, frame resonance, and participant mobilization', *International Social Movement Research*, vol. 1, pp. 197–217.

Snow, D. A., Rochford, B.Jr., Worden, S. K. & Benford, R. D. (1986) 'Frame alignment processes, micromobilization, and movement participation', *American Sociological Review*, vol. 51, pp. 464–481.

Tarrow, S. (2011) *Power in Movement: Social Movements in Contentious Politics* , 3rd edn, Cambridge University Press, New York.

Tilly, C. (2004) *Social Movements, 1768–2004*, Paradigm, Boulder, CO.

Tilly, C. (2006) 'WUNC', in *Crowds*, eds J. T. Schnapp & M. Tiews, Stanford University Press, Stanford, pp. 289–306.

Vinocur, N. & Barkin, N. (2009) 'G20 marches begin week of protests in Europe', *Reuters*. 28 March, [Online] Available at: http://www.reuters.com/article/2009/03/28/us-g20-britain-march-idUSTRE52R0TP20090328 (9 July 2011).

Weller, B. (2010) 'G20 protests in Seoul', *Demotix*, [Online] Available at: http://www.demotix.com/photo/504262/g20-protests-seoul (9 July 2011).

W. Lance Bennett is Professor of Political Science and Ruddick C. Lawrence Professor of Communication at University of Washington Seattle, where he directs the Center for Communication and Civic Engagement (www.engagedcitizen.org). *Address*: Department of Political Science and Communication, University of Washington, 101 Gowen Hall, Box 353530, Seattle, WA 98195, USA. [email: lbennett@uw.edu]

Alexandra Segerberg is a Research Fellow at the Department of Political Science, Stockholm University. *Address*: Department of Political Science, Stockholm University, 106 91 Stockholm, Sweden. [email: alex.segerberg@statsvet.su.se]

Received 15 September 2013
Revised 20 December 2013
19 February 2014
27 February 2014
Accepted 11 March 2014

Programmed method: developing a toolset for capturing and analyzing tweets

Erik Borra and Bernhard Rieder
Department of Media Studies, University of Amsterdam, Amsterdam, The Netherlands

Abstract

Purpose – The purpose of this paper is to introduce Digital Methods Initiative Twitter Capture and Analysis Toolset, a toolset for capturing and analyzing Twitter data. Instead of just presenting a technical paper detailing the system, however, the authors argue that the type of data used for, as well as the methods encoded in, computational systems have epistemological repercussions for research. The authors thus aim at situating the development of the toolset in relation to methodological debates in the social sciences and humanities.

Design/methodology/approach – The authors review the possibilities and limitations of existing approaches to capture and analyze Twitter data in order to address the various ways in which computational systems frame research. The authors then introduce the open-source toolset and put forward an approach that embraces methodological diversity and epistemological plurality.

Findings – The authors find that design decisions and more general methodological reasoning can and should go hand in hand when building tools for computational social science or digital humanities.

Practical implications – Besides methodological transparency, the software provides robust and reproducible data capture and analysis, and interlinks with existing analytical software. Epistemic plurality is emphasized by taking into account how Twitter structures information, by allowing for a number of different sampling techniques, by enabling a variety of analytical approaches or paradigms, and by facilitating work at the micro, meso, and macro levels.

Originality/value – The paper opens up critical debate by connecting tool design to fundamental interrogations of methodology and its repercussions for the production of knowledge. The design of the software is inspired by exchanges and debates with scholars from a variety of disciplines and the attempt to propose a flexible and extensible tool that accommodates a wide array of methodological approaches is directly motivated by the desire to keep computational work open for various epistemic sensibilities.

Keywords Twitter, Computational social science, Data collection, Analysis, Digital humanities, Digital methods

Paper type Conceptual paper

1. Introduction

The relatively recent flourishing of computer-supported approaches to the study of social and cultural phenomena – digital methods (Rogers, 2013), computational social science (Lazer *et al.*, 2009), digital humanities (Kirschenbaum, 2010), each with their set of significant precursors – has led to an encounter between technology and methodology that deeply affects the status and practice of research in the social

The authors would like to thank the special issue editors and the anonymous reviewers, as well as Noortje Marres, David Moats, Richard Rogers, Natalia Sanchez Querubin, Emma Uprichard, Lonneke van der Velden, and Esther Weltevrede for their useful comments, and Emile den Tex for his technical contributions. This project has benefited from a grant by the ESRC Digital Social Research Programme and the Centre for the Study of Invention and Social Process (Goldsmiths, University of London).

Aslib Journal of Information Management
Vol. 66 No. 3, 2014
pp. 262-278
© Emerald Group Publishing Limited
2050-3806
DOI 10.1108/AJIM-09-2013-0094

sciences and humanities. Statistics, modeling, and other formal methods have introduced strong elements of technicality long ago. But the study of very large sets of highly dynamic data, which, unlike, e.g. surveys, are not explicitly produced for scientific study, institutes computing as a methodological mediator (Latour, 2005), and brings along ideas, artifacts, practices, and logistics tied to the technological in a more far-reaching and radical fashion. Packages like SPSS have enabled, broadened, and standardized the use of computers and software in social research since the 1960s (Uprichard *et al.*, 2008). The recent explosion in research employing data analysis techniques, often focussed on social media and other online phenomena, however, propels questions of toolmaking – software design, implementation, maintenance, etc. – into the center of methodological debates and practices.

A number of commentators (Boyd and Crawford, 2012; Rieder and Röhle, 2012; Puschmann and Burgess, 2014) have called attention to the issues arising from the use of software to study data extracted from (mostly proprietary) software platforms that enable and orchestrate expressions and interactions of sometimes hundreds of millions of users. These issues include the methodological, epistemological, logistical, legal, ethical, and political dimensions of what is increasingly referred to as "big data" research. While such critical interrogation is necessary and productive, in this paper we take a different approach to some of the issues raised, by introducing and discussing an open-source, freely available data capture, and analysis platform for the Twitter micro blogging service, the Digital Methods Initiative Twitter Capture and Analysis Toolset (DMI-TCAT)[1]. Although we do not envision this research software to be a "solution" to the many questions at hand, it encapsulates a number of propositions and commitments that are indeed programmatic beyond the mere technicalities at hand. A presentation of such a tool cannot leave technical matters aside, but in this paper we attempt to productively link them to some of the broader repercussions of software-based research of social and cultural phenomena.

Although Cioffi-Revilla's assessment that "computational social science is an instrument-enabled scientific discipline, in this respect scientifically similar to microbiology, radio astronomy, or nanoscience" (Cioffi-Revilla, 2010, p. 260) needs to be nuanced, the argument that "it is the instrument of investigation that drives the development of theory and understanding" (Cioffi-Revilla, 2010) is not easy to dismiss when looking at research dealing with Twitter. The large number and wide variety of computational approaches, their status as mostly experimental tools, their application in disciplines often unaccustomed to computational principles, the pervasiveness of social media, and the speed of technological change – all these elements require us to pay much more attention to our instruments than we have been accustomed to. Having built such an instrument, we feel obliged to go beyond the presentation of architecture or results and account for the way we think – or hope – that our tool "drives" research in a more substantial way than solely solving particular technical and logistical problems. This desire possibly betrays our disciplinary affiliation. Media studies, in particular in its humanities bent, has long focussed on analyzing technologies as media, that is, as artifacts or institutions that do not merely transport information, but, by affecting the scale, speed, form, in short, the character of expression and interaction, contribute to how societies and cultures assemble, operate, and produce knowledge. Just as Winner (1980) pointed out that tools have politics too, we consider a research toolset such as DMI-TCAT to have epistemic orientations that have repercussions for the production of academic knowledge. Rather than glossing over them, we want to bring them to the front.

With these elements in mind, our paper proceeds in two distinct steps:

- We briefly summarize existing tools and approaches for Twitter analysis, discuss how they relate to academic research, and develop a set of guidelines or principles for our own contribution along the way.
- We present the design and architecture of DMI-TCAT, show how it addresses the concerns raised, and detail the analytical possibilities for Twitter research it provides. Through all of this, the relationship between toolmaking and methodology remains in focus.

2. Existing work

When highlighting the emerging antagonism between "Big Data rich" and "Big Data poor," Boyd and Crawford (2012) cite Twitter researcher Jimmy Lin as discouraging "researchers from pursuing lines of inquiry that internal Twitter researchers could do better." This quote echoes – at least if taken out of context – Savage and Burrows' (2007) diagnosis of a "coming crisis in empirical sociology": a marginalization of academic empirical work due to the ever increasing capacity and inclination of "knowing capitalism" (Thrift, 2005) to collect large amounts of data and to deploy a variety of methods to analyze them. However, instead of advocating retreat into the realms of synthesizing theory, they call for "greater reflection on how sociologists can best relate to the proliferation of social data gathered by others" (Savage and Burrows, 2007, p. 895) and for renewed involvement with the "politics of method" (p. 895) in both academic and private research. Rather than leaving areas like social media research to in-house scientists and marketers, we should be "critically engaging with the extensive data sources which now exist, and not least, campaigning for access to such data where they are currently private" (p. 896). We could not agree more with this assessment and would like to emphasize that the crisis Savage and Burrows diagnose goes beyond the question of access to data. The proliferation of actors involved in the analysis of online data – private and academic, coming from a wide variety of disciplines – has led to the formation of an epistemological battlefield where different paradigms, methods, styles, and objectives struggle for interpretive agency, i.e. for the power to produce (empirical) accounts of the ever expanding online domain. To be clear: the various technical, legal, logistical, and even ethical stumbling blocks for data analysis, and the ways in which the various actors are able or decide to react to them, have very real consequences for the actual knowledge produced and circulated.

In order to situate our own contribution and to develop a number of guiding principles we need to provide a short overview of existing strategies in Twitter research and discuss their limitations. The rough groupings we make, which revolve around logistical questions, precede concrete research designs and rather define a particular methodological space in which such concrete designs are then formulated.

Twitter's in-house research projects, or projects with cooperation agreements, have direct access to the full Twitter archive and are in the luxurious position to not have to worry about access to data, data completeness, or technical limitations – although legal and ethical considerations linger. At the same time, they are utterly dependent on the good will of the company. Their academic independence in terms of subject focus is doubtful, the tools and techniques used are often proprietary and can thus not be scrutinized, and only few projects will actually be selected in the first place.

Projects acquiring data through resellers such as DataSift or Gnip also gain access to the full archive of tweets and their metadata. Cost, however, is the main limiting factor to this approach: the pricy subscription models to those services are out of reach for small- and mid-sized research groups. As Twitter donated its entire archive to the US Library of Congress, a viable and cheap alternative may become available in the future. Any project working with data sourced from these archives, however, will have to rely on custom programming for analysis.

This brings us to online analytics platforms, which provide simple interfaces for both data acquisition and analysis, and are oriented toward either academic (e.g. DiscoverText[2], Truthy[3]) or commercial research (e.g. Topsy[4], Twitonomy[5], Hootsuite[6]). Those who interface with data resellers, such as DiscoverText, are again costly, but services collecting data through Twitter's Application Programming Interfaces (APIs) are often available for free or at reduced cost. These platforms and their dashboard-like interfaces can be very practical and useful, but are generally limited in terms of their analytical capacities, cannot be easily extended, and allow for little or no data export to stand-alone analytics software. Most problematic is the fact that they blackbox a large part of the research chain and generally follow a particular paradigmatic orientation. We do not think that commercial platforms should be dismissed outright, but it is clear that they are mainly focussing on the requirements of marketing professionals, emphasizing lists of "top" or "influential" users and content items. More academic platforms equally subscribe to specific paradigmatic approaches coming with prior assumptions about both data, e.g. Truthy considering spam as noise (McKelvey and Menczer, 2013), and method, e.g. DiscoverText focussing on the classification of tweets and Truthy on information diffusion. As such, researchers are restricted to their premises and analytical techniques.

If all that is needed is a set of tweets matching certain keywords, e.g. all the tweets containing a hashtag for a specific event, ad hoc or project-based custom capturing tools such as ScraperWiki[7], Google spreadsheets, or streamR[8] are commonly used. Just like custom programming, this approach affords flexibility, transparency, and control, but results may be difficult to verify or reproduce, bugs can occur, and significant technical skill needs to be acquired.

Two well-known examples of open-source capturing software, an approach that retains transparency and (some) flexibility and control while reducing the need for technical expertise, are 140kit[9] and TwapperKeeper. Both started out as public online services to capture, export, and – in the case of 140kit – analyze tweets, but had to close down when Twitter changed its terms of service in 2011. The source code for both projects was published online and yourTwapperKeeper[10] (yTK) in particular has been used by many humanities and social science scholars to capture tweets (see, e.g. Bruns and Liang, 2012). To facilitate and standardize research with yTK, which comes without built-in analytics, Bruns and Burgess (2012) published a set of useful GAWK scripts and we therefore initially used yTK to capture tweets. But the less technically inclined humanities and social science scholars we often work with found the GAWK scripts too difficult to handle. Our attempt to build a simpler analytics platform on top of yTK proved difficult: its database structure is not designed for fast analysis and omits many fields returned by the API[11]; its codebase is not updated on a regular basis; data is not stored as UTF-8 and languages using non-Latin character sets thus cannot be analyzed. Finally, we not only wanted to capture and analyze keyword based samples of tweets but also user timelines, 1 percent samples, follower networks, and other types of data available through Twitter's API.

Reviewing the possibilities and limitations of existing tools led us to the decision to build our own capture and analysis platform from the ground up. It also allowed us to develop a set of guiding principles that translate into a series of decisions or commitments on three interrelated levels. Concerning logistics, we attempt to lower the barrier of entry to Twitter research by providing a freely available platform built on publicly available data which requires little or no custom programming and scales to data sets of hundreds of millions of tweets using consumer hardware. Regarding epistemology, our tool emphasizes epistemic plurality by staying close to the units defined by the Twitter platform instead of storing aggregates, by allowing for a number of different sampling techniques, by enabling a variety of analytical approaches or paradigms, and by facilitating work at the micro, meso, and macro levels. On the level of methodology, finally, we provide robust and reproducible data capture and analysis, allow easy import and export of data, interlink with existing analytics software, and guarantee methodological transparency by publishing the source code.

In the next section, we provide a more detailed description of our system and show how these guiding principles have been translated into concrete design decisions.

3. DMI-TCAT

In line with the general architecture of DMI-TCAT, this presentation is divided into data capture, the way data are retrieved, enriched, and stored in a database, and data analysis, which includes all analytical operations that can be performed on the stored elements. While these two aspects have been developed in tandem, they are mostly independent: it is possible to use the toolset to only capture data, e.g. as alternative to yTK, or to only analyze them, e.g. by importing a data set captured with yTK. Figure 1 provides a basic overview of the system.

DMI-TCAT is written in PHP and organized around a MySQL database positioned between the capture and analysis parts of the system. Data are retrieved by different modules controlled in regular intervals by a supervisor script (using the cron scheduler present in all Unix-like operating systems), which checks whether the capturing processes are running and, if necessary, restarts them. A separate script translates shortened URLs. Database contents are analyzed in a two-stage process: the selection of a subsample precedes the application of various analytical techniques. In the following section, the various techniques for data capture are discussed in more depth.

3.1 Data capture

Sampling, i.e. the selection of items from a population of cases or elements, is a central concern when using online data. Because we are essentially dealing with data stored in

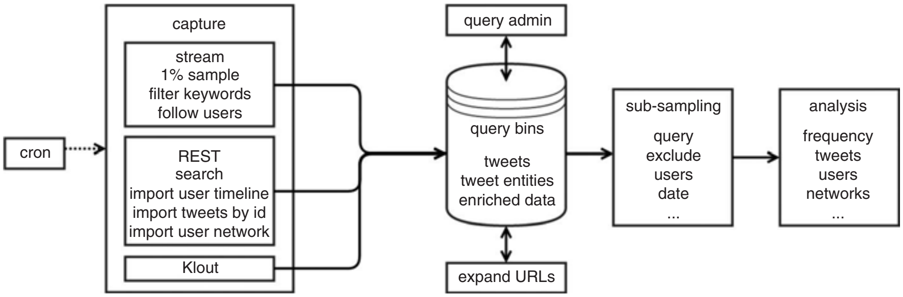


Figure 1.
Schema of the general architecture of DMI-TCAT

information systems, in this case Twitter's database, the properties of these systems determine, to a large extent, selection and retrieval possibilities. More fundamentally, they imply that the sociality they enable is structured and formalized: platforms like Twitter define basic entities (tweets, users, lists, hashtags, etc.), their characteristics (a tweet is no longer than 140 characters, a user account is associated with an image, etc.), and possible actions (a tweet can be retweeted, a user followed, etc.). The information system therefore goes a long way in framing what Uprichard calls "the ontology of the case" (Uprichard, 2013), simply by defining which entities can appear as a case in the first place and subsequently become part of a sample. As noted earlier, researchers in media studies have long recognized that media itself affect the character of expression and interaction passing through it. While, e.g. Facebook provides a formal and functional definition of "group," Twitter does not. That does not mean that a research design cannot operationalize such a construct by other means, e.g. by collecting accounts of a predefined group like members of a parliament, but because the functional characteristics of the Twitter platform mold actual use practices, the technical structuring of potential units of analysis is highly relevant. Focussing on Web media, digital methods (Rogers, 2013) thus urge to pay attention to the way in which digital objects are defined by and processed through online devices.

As data can be captured and stored in different ways the decisions made on this level have repercussions for analytical possibilities further down the chain. Hence, capturing tools already participate in the framing of the empirical as such. Attempting to facilitate epistemological and methodological diversity, DMI-TCAT closely follows Twitter's specific information structures[12], leaving the "primary" material untouched, while allowing for plasticity in sample design and easy ways to create subsamples from captured data sets.

Apart from technical specifications, Twitter also defines and regulates the modes and scope of access to any data (Puschmann and Burgess, 2014). Legal constraints, API definitions, rate limits for query calls, whitelisting, and data sharing agreements are among the many possibilities the company has to design the ways its data can become part of a research project. Because APIs are designed to enhance Twitter's value as a commercial platform by allowing third-party developers to build applications on top of it, the needs of researchers are not explicitly taken into account. Toolsets like DMI-TCAT thus repurpose these technical interfaces for research. The following section shows by which different technical pathways data enter into the system.

3.1.1 Data acquisition. DMI-TCAT relies on Twitter's APIs and is therefore bound to their possibilities and limitations. While we do not require familiarity with these technical interfaces, we notify users when problems occur, e.g. when rate limits are exceeded. Our tool connects to Twitter using the tmhOAuth[13] library and retrieves tweets via both the streaming API and the REST API[14]. We use the former for three different sampling techniques. First, researchers can capture a "1 percent" random sample[15] of all tweets passing through Twitter, which can then be used for macro- and meso-level investigations and for baselining samples retrieved by other means (Gerlitz and Rieder, 2013). Second, we use the statuses/filter endpoint[16] to "track" tweets containing specific keywords in real-time, which is probably the most common way to create a sample of tweets. To give researchers maximum flexibility and specificity, a collection is defined in a so-called "query bin," i.e. a list of tracking criteria[17] consisting of single or multiple keyword queries, hashtags, and specific phrases. For example, a bin like (globalwarming, "global warming," #IPCC) would retrieve all tweets containing one of these three query elements and combine them into a single data set, stored as a group of related tables in the database. Third, our system allows for following tweets from a specified set of up to 5,000 users. This is particularly interesting when studying a set of manually selected accounts, such as members of a parliament or other expert lists.

One of the main limits of the streaming API is that it cannot provide historical tweets. The REST API[18], however, enables a search for tweets up to about a week old and although it explicitly omits an unknown percentage of tweets[19], a data set can be started by retrieving tweets up to about a week old. While this is far from ideal, it might be the only feasible way to record traces of an unanticipated event. In the same spirit, we use search to fill gaps in data capture resulting from network outages or other technical problems. Finally, the REST API allows for retrieving the last 3,200 tweets for each user in a set, providing a level of historicity for user samples, and the retrieval of follower/followee networks.

Additionally, DMI-TCAT takes Twitter's data sharing policy into account, which allows for sharing of tweet ids but not of messages and metadata themselves[20]. Our tool is therefore able to reconstruct a data set from a list of ids and can export such a list as well. Along the same line, we provide import scripts for yTK databases or a set of Twitter JSON files captured by other means (e.g. streamR).

Taken together, these possibilities allow for a wide array of sampling techniques, in line with the principle of methodological flexibility.

3.1.2 Data storage and performance. Following the arguments for paying close attention to Twitter's informational structures outlined above, our database layout mimics the shape of the data returned by the API[21]. This means that tweets and their metadata, hashtags, URLs, and mentions are stored in separate tables. DMI-TCAT therefore does not need to extract those entities anymore, making querying the database much faster.

While there are indeed limits to the amount of data one can store and analyze without moving into the complicated and costly world of distributed computing, a well-designed and indexed database structure, combined with optimized database queries, means that off-the-shelf consumer hardware can handle much larger quantities of tweets than sometimes argued (Bruns and Liang, 2012). We are currently running DMI-TCAT on a cheap Linux machine with four processor cores, a 512GB SSD, and 32GB of RAM, using the default LAMP[22] stack. At the time of writing, we have captured over 700 million tweets and basic analyses for even the largest query bins – over 50 million tweets in a single data set – generally complete in under a minute, allowing for iterative approaches to analysis. More complex forms of analysis, such as the creation of mention networks, can take several minutes to complete, however. While we have not systematically evaluated how far our architecture can scale, it seems safe to say that hundreds of millions of tweets in a single data set should still be workable, but moving on to the next order of magnitude would certainly require a fully distributed approach to tools and infrastructure that is beyond the scope of our software.

3.1.3 Data enrichment. One area where our system strays from simply capturing and storing data provided by the API is data enrichment. We currently follow two directions: URL expansion and the addition of Klout scores. First, many URLs passing through Twitter are shortened, and although Twitter provides the "final" URL for its own shortening service, this is not the case for third party shortening services such as bit.ly. However, in tradition with other digital methods tools (Rogers, 2010), URLs and, in particular, domain names are considered as crucial components of a tweet's message and a robust means for actor identification and content qualification. DMI-TCAT therefore includes a script that follows all URLs to their endpoint, adds the location to the URL table, and extracts the domain name. Second, we provide the option to retrieve users' Klout score[23], a proprietary metric set in the sociometric tradition that produces an "influence" rating based on data from eight different social media platforms. While caution is in order when using proprietary metrics, Klout scores are commonly used and offer a glimpse into users' activities beyond the Twitter platform.

3.2 Data analysis

In contrast to yTK, DMI-TCAT is not limited to capturing data but also provides analytical techniques to researchers, in a way that strikes a balance between ease of use and analytical flexibility. We try to enable approaches spanning the "three major areas of analysis" (Bruns and Liang, 2012) in Twitter research – tweet statistics and activity metrics, network analysis, and content analysis – but also facilitate geographical analysis, ethnographic research, and even textual hermeneutics. Our tool can thus be used in a wide variety of projects, including studies of everyday conversation, breaking news, crisis communication, political activism (citizen), journalism, second screen applications, lifestyle and brand communication, information diffusion, social patterns, ideological frames, sentiment analysis, prediction, and so forth.

Besides providing a variety of analytical pathways and facilitating the integration of additional modules, we emphasize epistemic plurality – and lower the cost of development – by embracing Marres' (2012) assessment that "social research becomes *noticeably* a distributed accomplishment: online platforms, users, devices, and informational practices actively contribute to the performance of digital social research" (Marres, 2012, p. 139). Pushing this "redistribution of method" further, we enable the export of derived data in standard formats to be analyzed in software packages, chosen by researchers themselves, over interactive interfaces and ready-made (visual) outputs. This makes the tool less convenient for users without experience in data analysis, but the recent emergence of tools such as the Gephi[24] graph visualization and manipulation software, which are at the same time easy to use and much more powerful than any Web interface, justifies this compromise.

This points to a conundrum that any toolmaker faces: to what extent and in what way does research software actually shape research practices? How to justify decisions and how to organize the design process? While there are few firm guidelines, the above describes a particular balance between methodological flexibility and ease of use that is the outcome of constant interaction with students of a large MA class on digital methods, with the participants of multiple workshops, and with several research projects at the University of Amsterdam relying on DMI-TCAT. To accommodate this interaction, we rely on an agile software development approach using rapid prototyping, iterative updating, and a modular architecture. While decisions necessarily have to be made, they can be shared in a flexible way because tool-making and actual research remain tightly coupled. Most of the analytical outputs DMI-TCAT provides were thus built in response to particular analytical requirements from either ourselves or from researchers we have been working with. However, the collaborative setting cannot fully alleviate the fact that toolmakers necessarily intervene deeply in the epistemic process of methods development, by framing ideas in terms of feasibility, cost, formalization, and so forth, but also by constantly translating and connecting social science and humanities concepts to computational techniques.

In the following sections, we will detail the analytical techniques implemented in DMI-TCAT, starting with sub-sampling from stored data and continuing with summary presentations of different analytical outputs.

3.2.1 Sub-sampling. DMI-TCAT enables the flexible constitution of a subsample. After selecting a data set, as defined by a query bin, different techniques to filter the data set are available.

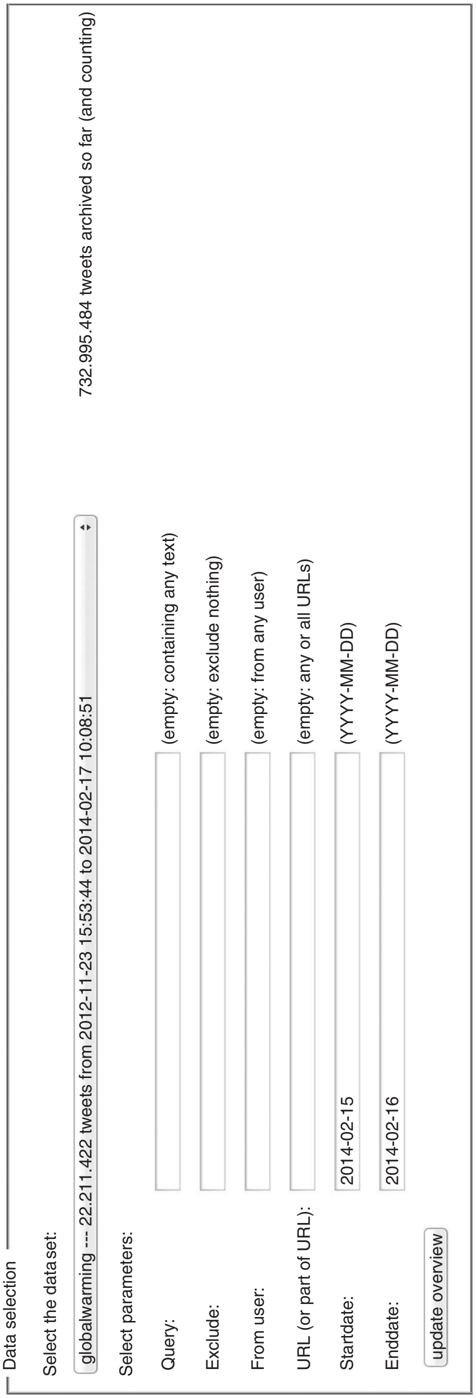
By sub-selecting from the data set, a user can zoom in on a specific time period or on tweets matching certain criteria (Figure 2). She can choose to include only those tweets matching a particular phrase such as a word, hashtag, or mention; she can exclude tweets matching a specific phrase; finally, she can focus on tweets by particular users or tweets mentioning a specific (part of a) URL. All input fields accept multiple phrases or keywords to specify (AND) or expand (OR) the selection via Boolean queries. After updating the overview, a summary of the selection is generated (Figure 3). While these filters are far from exhaustive, they allow for both the constitution of a subsample and what could be described as "interactive probing," i.e. the back and forth movement between query and overview that "progresses in an iterative process of view creation, exploration, and refinement" (Heer and Shneiderman, 2012). This also echoes Uprichard's argument that social research rarely proceeds in linear fashion, but that "cases are 'made', both conceptually and empirically, by constantly and iteratively re-shaping and re-matching theory and empirical evidence together" (Uprichard, 2013, p. 5).

In addition, sub-sampling can be seen as a means for non-destructive data cleaning in the sense that tweets matching specific criteria can be excluded without being deleted. While this is currently implemented in rudimentary fashion only, the question of data cleaning is crucial for both the reliability of the data and the question of epistemic plurality, or as Gitelman (2013, p. 5) puts it "data […] need to be understood […] according to the uses which they are and can be put"; one researcher's noise is another one's object of study.

The overview interface (Figure 3) lists the current selection criteria and shows the number of tweets in the subset, the number of distinct users, and the proportion of tweets containing URLs. Additionally, a line graph shows the frequency of tweets, distinct users, distinct locations, and geo-tagged tweets per hour or per day, depending on the scope of the selection. If a search query is specified, a second line graph indicates the relationship between the subset and the full dataset. The example in Figure 3 thus not only shows the shrinking absolute number of tweets mentioning (snowden) in our "prism" dataset over the month of July 2013, but also the relative decline of tweets mentioning the whistleblower's name in relation to the whole data set.

Subsequent analytical techniques apply to the selected subsample, although the full data set is, indeed, a possible selection. With the exception of several interactive modules, all analyses are provided as exports in standard tabulated formats or as network files. Filenames include filters and settings used in the interface so that researchers know how data has been derived and which software version was used. The following four sections describe the various types of techniques currently implemented in our tool and mimic the sections of the actual interface.

3.2.2 Tweet statistics and activity metrics. The first set of exports covers some of the basic statistics of the sub-sample one may want to consult. To get a quick characterization of the types of tweets in the sample a table is provided with the total



Data selection

Select the dataset:

globalwarming --- 22.211.422 tweets from 2012-11-23 15:53:44 to 2014-02-17 10:08:51

732.995.484 tweets archived so far (and counting)

Select parameters:

Query: (empty: containing any text)

Exclude: (empty: exclude nothing)

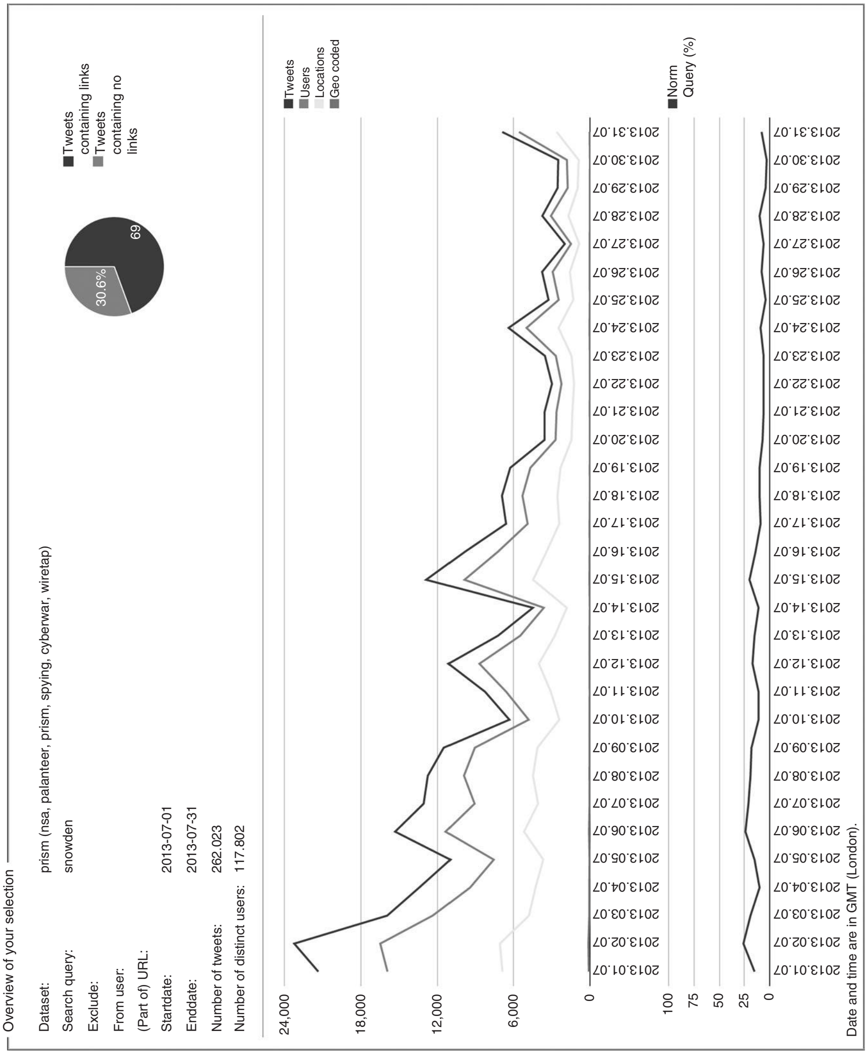
From user: (empty: from any user)

URL (or part of URL): (empty: any or all URLs)

Startdate: 2014-02-15 (YYYY-MM-DD)

Enddate: 2014-02-16 (YYYY-MM-DD)

Figure 2.
Screenshot of the "data selection" part of the DMI-TCAT interface



Overview of your selection

Dataset: prism (nsa, palanteer, prism, spying, cyberwar, wiretap)

Search query: snowden

Exclude:

From user:

(Part of) URL:

Startdate: 2013-07-01

Enddate: 2013-07-31

Number of tweets: 262.023

Number of distinct users: 117.802

Date and time are in GMT (London).

Figure 3.
Screenshot of the "overview" section of the DMI-TCAT interface

number of tweets, the number of tweets with URLs, hashtags, and mentions, as well as the number of retweets and the number of unique users in the selection. To characterize user activity and visibility, a table is provided which lists the minimum, maximum, average, and median number of tweets sent, the number of users mentioned, the number of followers and followees, and the number of URLs tweeted, all per user. Furthermore, we follow emerging standards in Twitter research and allow for easy analysis of basic platform elements over time (cf. Bruns and Burgess, 2012). This includes counts of hashtags, user tweets and mentions, URLs and domain names, as well as retweets[25].

Each of these outputs can be segmented into hourly, daily, weekly, monthly, and yearly intervals; self-chosen intervals – e.g. to delineate distinct periods – are possible as well and permit fine-grained temporal analysis. Although a full discussion of the various metrics and their uses has to be deferred to a future publication, it is important to mention that certain outputs provide deeper analytical perspectives: the hashtag-user output, for example, not only provides a hashtag count per time interval, but also the number of distinct users sending tweets containing the hashtag, the number of distinct users mentioned in tweets with the hashtag, and the number of tweets mentioning the hashtag.

3.2.3 Tweet exports. This section of the interface regroups modules providing lists of actual tweets for further analysis. A random set of a user-specified number of tweets can facilitate content analysis by providing a sample of items to be (manually) coded into categories or otherwise analyzed. It is also possible to simply export all tweets and their metadata from the current selection or, alternatively, only those that have been retweeted or come with geo-location data.

A statistical exploration of the data via the methods outlined in the previous section, combined with different analyses of actual tweets enables powerful mixed methods approaches (cf. Lewis *et al.*, 2013). For example, researchers can export a chronological list of the most retweeted messages, which is an interesting means to reconstruct and narrate the timeline of an event (Rogers *et al.*, 2009). From close reading to text mining, the easy availability of actual tweets is crucial for both quantitative and qualitative examinations of content.

3.2.4 Networks. The third set of outputs focusses on network perspectives and produces outputs in either GEXF or GDF formats. The difference with statistical approaches lies not so much in what is being looked at, but rather in how the data are represented and analyzed. At the moment, the main focus lies on users, hashtags, and URLs – and the various relationships between these entities. Two outputs represent interaction networks where users are connected either through mentions or through direct replies. Because these files can be opened in different graph analysis tools, a wide variety of social network analysis techniques can be applied. This affords perspectives on interaction patterns that go beyond mere frequency and allows, for example, identifying cliques or sub-conversations.

A co-hashtag network output allows for a type of content analysis that focusses on relationships between these signal words: if two hashtags appear in the same tweet, a link is established; the more often they co-occur, the stronger the link. By applying network analysis techniques, one can get an overview of the subject variety in a set of tweets and analyze relationships between subtopics.

Finally, there are a number of bipartite graphs outputs – networks containing entities of two kinds constituted through co-occurrence in a tweet – in particular hashtag-user and hashtag-URL/domain networks. Figure 4 provides a short example

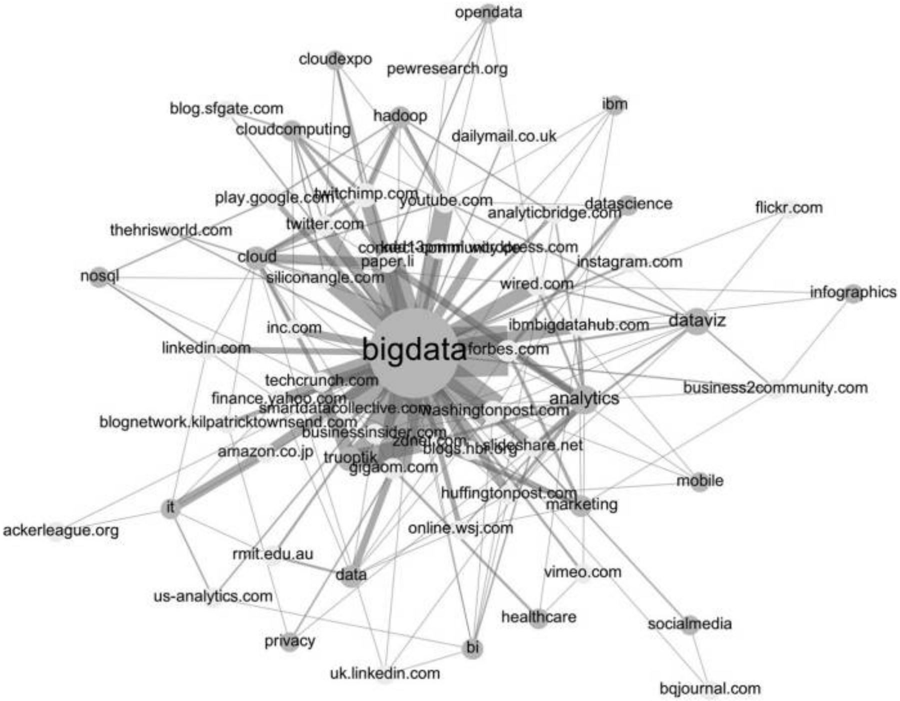


Figure 4.
Gephi graph visualization of a bipartite graph from our "datascience" data set: hashtags are in dark gray and domain names in light gray

for the latter. These techniques allow for the structural analysis of relationships between entities and are particularly useful for locating actors (users or domains) in relation to issues (hashtags).

3.2.5 Experimental modules. Because DMI-TCAT is modular, it is easy to add new analytical techniques. A series of experimental modules provide interactive interfaces or dashboards rather than file exports. A detailed presentation is beyond the scope of this paper, but our "cascade" module, a means to visually explore temporal structures and retweet patterns, serves as an example.

This module (Figure 5) provides a ground-level view of tweet activity by either charting every single tweet in the current selection or only those above a certain retweet threshold. User accounts are distributed vertically; tweets – shown as dots – are spread out horizontally over time. Lines indicate retweets. At the top we see the typical activity pattern of a retweet bot (line of dark dots). This view requires a large screen and is limited to small data selections, but because tweet text becomes visible when hovering over a node, it allows for the close reading of a conversation or debate and, in a sense, links to ethnographic observation.

4. Conclusions

In this paper, we have described a tool to capture and analyze data from Twitter. We have shown how particular design decisions can be related to wider considerations concerning the role of software in academic research. Our proposition is not simply a "solution" to a set of "problems." Rather, it is an attempt to connect the question of toolmaking for social and cultural research to debates dealing with the "politics of method" (Savage and Burrows, 2007, p. 895) in ways that are not merely theoretical or

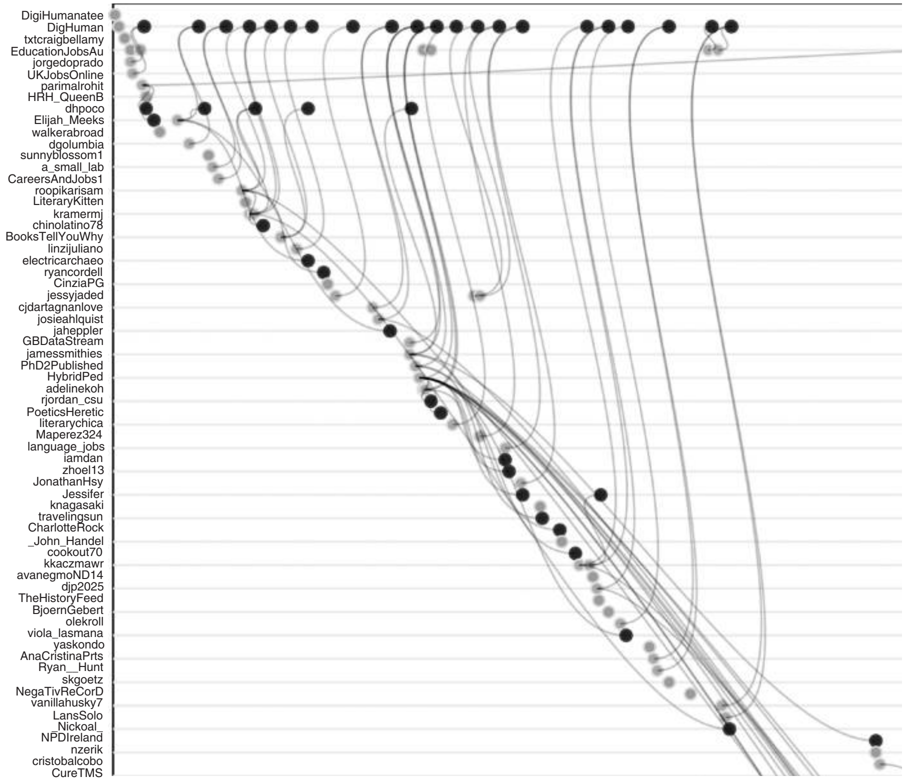


Figure 5.
Partial screenshot of the "cascade" module interface

critical. Platforms like Twitter pose a number of fundamental challenges to scholars. Beyond being attentive to these questions, we have to ask how the tools we use to practice research can proactively take those challenges into account. The canonical style of both research reporting and technical publication leave little space to connect to fundamental interrogations of methodology and its repercussions for the production of knowledge. But the very nature of computational methods, which deeply entangle research design with technical work, requires us to engage toolmaking from different angles. When even small decisions in database design can lead to huge differences in performance, potentially having profound effects on the way researchers interact with the tools and data, we realize that even details in implementation can have substantial epistemic effects. Are we missing a genre of academic text that permits a combination of technical presentation and general methodological discussion? The direct relationship between engineering questions and methodological considerations is a subject that is often neglected and merits much more critical debate.

This paper in no way suggests to serve as a blueprint for such an endeavor, but in order to engage the enormous methodological challenges, we feel compelled to experiment with forms of writing and academic expression that attempt to span various disciplinary traditions, even if this might lead to disorientation and friction with established conventions.

The design of DMI-TCAT is inspired by exchanges and debates with scholars from a variety of disciplines and our attempt to propose a flexible and extensible tool that accommodates a wide array of methodological approaches is directly motivated by the desire to keep computational work open for various epistemic sensibilities.

Notes

1. Available at https://github.com/digitalmethodsinitiative/dmi-tcat (accessed February 19, 2013).
2. http://discovertext.com (accessed September 14, 2013).
3. http://truthy.indiana.edu (accessed September 14, 2013).
4. http://topsy.com (accessed September 14, 2013).
5. http://twitonomy.com (accessed September 14, 2013).
6. https://hootsuite.com (accessed September 14, 2013).
7. https://scraperwiki.com (accessed September 14, 2013).
8. http://cran.r-project.org/web/packages/streamR/ (accessed September 14, 2013).
9. https://github.com/WebEcologyProject/140kit (accessed September 14, 2013).
10. https://github.com/540co/yourTwapperKeeper (accessed September 14, 2013).
11. DMI-TCAT stores every field returned by Twitter (around 40 per tweet – if the tweet contains mentions, hashtags, and URLs) while yTK only stores 13 of the most basic fields per tweet and excludes fields such as retweet_id, in_reply_to_status_id, entities, and many fields related to the sender of the tweet.
12. The API documentation is thus an essential part of DMI-TCAT's documentation. https://dev.twitter.com/docs/platform-objects (accessed September 1, 2013) specifies Twitter's entities and possible actions.
13. tmhOAuth by Matt Harris is available at https://github.com/themattharris/tmhOAuth. It implements all possible calls to the Twitter APIs in PHP (accessed September 1, 2013).
14. For an explanation of the differences, see https://dev.twitter.com/docs/streaming-apis (accessed September 12, 2013).
15. https://dev.twitter.com/docs/api/1.1/get/statuses/sample (accessed September 12, 2013).
16. https://dev.twitter.com/docs/api/1.1/post/statuses/filter (accessed September 12, 2013).
17. Tracking criteria follow https://dev.twitter.com/docs/streaming-apis/parameters#track (accessed September 12, 2013).
18. https://dev.twitter.com/docs/api/1.1 (accessed September 12, 2013).
19. "The Search API is not complete index [sic] of all Tweets, but instead an index of recent Tweets. At the moment that index includes between 6-9 days of Tweets." https://dev.twitter.com/docs/using-search (accessed September 12, 2013).
20. According to I.4.A from https://dev.twitter.com/terms/api-terms (accessed September 10, 2013), "If you provide downloadable datasets of Twitter Content or an API that returns Twitter Content, you may only return IDs (including tweet IDs and user IDs)."
21. We aim to always incorporate all fields returned by Twitter's APIs. We store all data in UTF-8 and are thus able to capture and analyze tweets in any language. See https://dev.twitter.com/docs/counting-characters (accessed September 10, 2013) for more information.

22. http://en.wikipedia.org/wiki/LAMP_%28software_bundle%29 (accessed September 12, 2013). DMI-TCAT has been tested on Linux and OSX.
23. http://klout.com/corp/how-it-works (accessed September 7, 2013).
24. Available as free software on http://gephi.org (accessed September 12, 2013).
25. We identify retweets by using Twitter's retweet_id API field as well as by grouping "identical" tweets, thus also including possible manual retweets.

References

Boyd, D. and Crawford, K. (2012), "Critical questions for big data", *Information, Communication & Society*, Vol. 15 No. 5, pp. 662-679.

Bruns, A. and Burgess, J. (2012), "Researching news discussion on Twitter: new methodologies", *Journalism Studies*, Vol. 13 Nos 5-6, pp. 801-814.

Bruns, A. and Liang, Y.E. (2012), "Tools and methods for capturing Twitter data during natural disasters", *First Monday*, Vol. 17 No. 4, p. 5, available at: http://firstmonday.org/ojs/index.php/fm/article/view/3937/3193 (accessed February 26, 2014).

Cioffi-Revilla, C. (2010), "Computational social science", *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 2 No. 3, pp. 259-271.

Gerlitz, C. and Rieder, B. (2013), "Mining one percent of twitter: collections, baselines, sampling", *M/C Journal*, Vol. 16 No. 2, available at: www.journal.media-culture.org.au/index.php/mcjournal/article/view/620 (accessed February 26, 2014).

Gitelman, L. (Ed.) (2013), *Raw Data' Is an Oxymoron*, MIT Press, Cambridge, MA.

Heer, J. and Shneiderman, B. (2012), "Interactive dynamics for visual analysis", *Queue*, Vol. 10 No. 2, pp. 30-55.

Kirschenbaum, M.G. (2010), "What is digital humanities and what's it doing in English departments?", *ADE Bulletin*, Vol. 150 No. 7, pp. 55-61.

Latour, B. (2005), *Reassembling the Social: An Introduction to Actor-Network-Theory*, Oxford University Press, New York, NY.

Lazer, D., Pentland, A.(S.), Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Van Alstyne, M. (2009), "Life in the network: the coming age of computational social science", *Science*, Vol. 323 No. 5915, pp. 721-723.

Lewis, S.C., Zamith, R. and Hermida, A. (2013), "Content analysis in an era of big data: a hybrid approach to computational and manual methods", *Journal of Broadcasting & Electronic Media*, Vol. 57 No. 1, pp. 34-52.

McKelvey, K. and Menczer, F. (2013), "Design and prototyping of a social media observatory", *Proceedings of the 22nd international conference on World Wide Web companion, Rio de Janeiro, International World Wide Web Conferences Steering Committee, Geneva*, pp. 1351-1358.

Marres, N. (2012), "The redistribution of methods: on intervention in digital social research, broadly conceived", *The Sociological Review*, Vol. 60 No. S1, pp. 139-165.

Puschmann, C. and Burgess, J. (2014), "The politics of twitter data", in Weller, K., et al. (Eds), *Twitter and Society*, Peter Lang Publishing, New York, NY, pp. 43-54.

Rieder, B. and Röhle, T. (2012), "Digital methods: five challenges", in Berry, D.M. (Ed.), *Understanding Digital Humanities*, Palgrave Macmillan, Basingstoke, pp. 67-84.

Rogers, R. (2010), "Mapping public web space with the Issuecrawler", in Brossard, C. and Reber, B. (Eds), *Digital Cognitive Technologies: Epistemology and Knowledge Society*, Wiley, London, pp. 115-126.

Rogers, R. (2013), *Digital Methods*, MIT Press, Cambridge, MA.

Rogers, R., Jansen, F., Stevenson, M. and Weltevrede, E. (2009), "Mapping democracy", in Finlay, A. (Ed.), *Global Information Society Watch 2009*, Association for Progressive Communications and Hivos, Uruguay, pp. 47-57, available at: www.giswatch.org/fr/node/158 (accessed February 26, 2014).

Savage, M. and Burrows, R. (2007), "The coming crisis of empirical sociology", *Sociology*, Vol. 41 No. 5, pp. 885-899.

Thrift, N. (2005), *Knowing Capitalism*, Sage, London.

Uprichard, E. (2013), "Sampling: bridging probability and non-probability designs", *International Journal of Social Research Methodology*, Vol. 16 No. 1, pp. 1-11.

Uprichard, E., Burrows, R. and Byrne, D. (2008), "SPSS as an "inscription device": from causality to description?", *The Sociological Review*, Vol. 56 No. 4, pp. 606-622.

Winner, L. (1980), "Do artifacts have politics?", *Daedalus*, Vol. 109 No. 1, pp. 121-136.

About the authors

Erik Borra is a PhD Candidate and Lecturer at the University of Amsterdam's MA Program in New Media. His research concerns the Web as a source of data for social and cultural research, paying particular attention to search engine queries, Wikipedia edit histories, and social networks. Erik is also scientific programmer for the Digital Methods Initiative and is currently involved in the European research project "Electronic Maps to Assist Public Science" (EMAPS). Erik Borra is the corresponding author and can be contacted at: borra@uva.nl

Dr Bernhard Rieder is an Associate Professor of New Media at the University of Amsterdam. Besides developing and theorizing digital methods, his research focusses on the history, theory, and politics of software, particularly on the role of algorithms in social processes and the production of knowledge. He has worked as a Web Programmer on various projects and is currently writing a book that investigates the history and cultural significance of information processing techniques.

SSI

Introduction: Digital methods for the exploration, analysis and mapping of e-diasporas

Social Science Information
51(4) 451–458
© The Author(s) 2012
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0539018412456918
ssi.sagepub.com
SAGE

Dana Diminescu
TelecomParisTech and Fondation Maison des Sciences de l'Homme, Paris

One of the major changes affecting diasporas the world over since the 1980s has been the increasing number of communities scattered throughout physical space, along with new forms of presence, regrouping, interaction and mobilization within digital territories.

This change calls for a renewal in epistemological approaches. The topics under study, as well as the conceptual and methodological tools used to analyse them, need to be reconsidered in the face of this evolution of diasporas. The articles published in this issue of SSI[1] bear witness to such an effort: researchers and engineers involved in the *e-Diasporas Atlas*[2] project have sought to find the most appropriate concepts, tools and methods to explore the Web of diasporas, based on a number of case studies. This work represents a vast new area of investigation, which is still under way.

In this introduction, we examine the different conceptual tools used during the research, analyse their relevance for the different diasporic communities on the Web and present the methodological chain developed within the *e-Diasporas Atlas* project as well as the most important findings.[3]

The concepts

In several articles, and specifically in another issue of SSI, we have shown the emergence of a new migrant figure: the connected migrant.[4] S/he is no longer defined solely by life-experiences of disruptions and antagonisms – which have constantly been upheld as the organizing principles of any theoretical reflection on the uprooted migrant and his 'twofold absence' – but by different forms of 'presence at a distance'.

This change, which we had initially studied at the fundamental level of the migrant him/herself, can also be observed at the collective level of diasporas and transnational networks.

What kinds of diasporas are formed by connected migrants? Do the networks and interactions of migrants scattered throughout the world, which we have been able to

Corresponding author:
Dana Diminescu, FMSH, 190 avenue de France, Paris 75013, France
Email: dana.diminescu@telecom-paristech.fr

observe and visualize through an exploration of their traces on the Web, reveal traditional or novel functions of diasporas? Are these ‘e-diasporas’ an extension of physical diasporas, or merely their mirror image? Are they the source of new diaspora communities? Or are they, instead, an echo-chamber of globalization – of a society which is itself a diaspora in the making? All these questions show how difficult it is to find a generally accepted definition of what an e-diaspora is. Discussions around this concept are not settled, even after many years of work. In fact, this publication aims at stimulating further debate.

Historically, the emergence of e-diasporas occurred along with the diffusion of the Internet and the development of multiple online public services. At the end of the 1990s, a number of institutions joined forces with the new ‘e’-technologies (e-administration, e-democracy, e-education, e-healthcare, e-culture, e-tourism), which gave rise to the first presence on the Web of associations run by migrant populations. If the earliest websites were those produced by IT professionals, we soon saw the diffusion of the Web into all the *diasporic* communities and at all levels within them. The past ten years have witnessed the use of both Webs 1.0 and 2.0 (blogs) in these communities as well as the widespread appropriation of the various social-networking platforms (Facebook, Twitter, LinkedIn, etc.).

What we call *e-diaspora* is a migrant collective that organizes itself and is active first and foremost on the Web: its practices are those of a community whose interactions are ‘enhanced’ by digital exchange. An e-diaspora is also a dispersed collective, a heterogeneous entity whose existence rests on the elaboration of a common direction, a direction not defined once and for all but which is constantly renegotiated as the collective evolves. An e-diaspora is an unstable collective because it is redrawn by every newcomer. It is self-defined, as it grows or diminishes not by inclusion or exclusion of members, but through a voluntary process of individuals joining or leaving the collective – simply by establishing hyperlinks or removing them from websites.

An e-diaspora is both ‘online’ and ‘offline’. We are therefore interested in both the digital ‘translations’ of ‘physical’ actors/phenomena (the online activities of associations, for example) and the specifically (‘natively’) digital actors/phenomena (e.g. a forum and its internal interactions), which are sometimes called pure players. The question of ‘rub-offs’ – reciprocal influence between these two sorts of Web entities – is of capital importance in analysing an e-diaspora. It is thus clear that the research carried out in the context of the *e-Diasporas Atlas* presupposes knowledge of the diaspora in question and also implies knowledge of the Web and an appreciation of the singularity of the exchanges that take place there.

We prefer the term ‘e-diaspora’ to that of ‘digital diaspora’ because the latter may lead to confusion, given the increasingly frequent use of the notions of ‘digital native’ and ‘digital immigrant’, in a ‘generational’ sense (distinguishing those born before from those born during/after the digital era). The object of the *e-Diasporas Atlas* is not this ‘digital migrant’, however, but the *connected migrant* in his/her social and institutional context.

An *e-diaspora corpus* is a list of websites. The constitution of a corpus of websites is the method used to ‘capture’ an e-diaspora. To start exploring the Web of diasporas, we first need to identify diasporic websites.

A ‘*diasporic*’ *website* or migrant website is a website created or managed by migrants and/or that deals with them (at any rate, a site for which migration or diasporas is a defining theme). This can be a personal site or blog, the site of an association, a portal/forum, an institutional site, or anything similar. Usage is not the criterion: a site often consulted by migrants (a media site, for example) is not necessarily a migrant site. What distinguishes ‘activity’ is first and foremost the production of content and the practice of citation (hyperlinks). On the other hand, a migrant site need not necessarily be located in a foreign country; it may just as easily be in the country of origin. Migrant sites testify to a given e-diaspora’s occupation of the Web.

Some researchers have built their collection by starting with websites they already knew from the fieldwork they had done on migrations (Ingrid Therwath, Houda Asal, Tristan Bruslé, Yann Scioldo Zürcher). Others have used keywords and Google (Priya Kumar, Anat Ben-David, Francesco Mazzucchelli) or Twitter (e.g. through the hashtag #right2vote, for Marta Severo and Eleonora Zuolo) before crawling the Web with the tools of the *e-Diasporas Atlas*. The choice and classification of websites relies entirely on the researcher’s expertise. Such choices have triggered debates which are still ongoing within the *e-Diasporas Atlas* community of researchers. This community is multidisciplinary, and the diversity of its members can be seen in the collection and categorization of websites. Houda Asal chose to work only on the websites of associations from the Lebanese diaspora. Ingrid Therwath chose to focus on websites related to the *hindutva* ideology. However, nearly all of them decided to collect official websites for a better understanding of the links between diasporas and their countries of origin. And all of them analysed the contents of each website, before deciding if they should keep it in their collection. The domain names, the languages used, the type of publication, the geolocation of the website, these are categories which have been used for every one of the case studies.

Frontier sites or neighbouring sites have also been collected.

A *neighbouring site* is a non-migrant site (or one belonging to an e-diaspora other than the one being studied) which distinguishes itself by its strong connection with the (migrant) sites of a given e-diaspora (governmental or media sites of the country of origin, for example). However, not every site strongly linked to an e-diaspora is necessarily a neighbouring site. To be one it needs to be ‘specific’ to the diaspora in question, which is why sites ‘on the fringes of’ the majority of Web communities, particularly those in the upper layers of the Web, Google, YouTube, Facebook and so on, are not counted as ‘neighbours’.

In the *e-Diasporas Atlas*, a list of neighbouring sites may be drawn up alongside that of migrant ones. These neighbouring sites discovered during the prospecting phase are not crawled during subsequent prospection but only during the phase of validation so as to gather together all links with the migrant site.

An ensemble of the ‘migrant sites’ and ‘neighbouring sites’ (see below) of a given diaspora – whether such sites are ‘living’ or ‘dead’ – constitutes the *web of diasporas*. In a sense, this can be understood as the web ecosystem of a diaspora.

e-Diasporas methodology

The digital methodological chain and the tools we developed for building the *e-Diasporas Atlas* aim at mapping and analysing the occupation of the Web by diasporas. The chain is composed of four interlocking steps: (1) equipped Web exploration and corpus building; (2) data enrichment (location, languages, text-mining); (3) network visualization-manipulation and graph interpretation; (4) collaborative sharing of (raw) data and findings.

Step 1: Web exploration

In order to compile a chapter of the Atlas, the first step is to build (and circumscribe) a *corpus* of websites. As we have already highlighted, the researcher plays a crucial role in this process inasmuch as his/her knowledge of the fieldwork allows him/her to select with discrimination the relevant resources for a given diaspora.

An e-diaspora is ‘captured’ by putting together a corpus of websites. This method entails breakdown and selection processes that allow a diaspora web to be extracted. But definition is also necessary because an e-diaspora presents itself to the researcher only as a product of this ‘excision’ performed upon the Web. Similarly, it is only because of such exploration/selection, the filtering/circumscription of a corpus, that what a migrant site actually is takes on meaning.

In order to complete this stage of *collection*, the researcher needs to be *equipped*. The identification of relevant websites is achieved semi-automatically thanks to a software called *Navicrawler*, which makes it possible to scan *web grounds* using a web-browser. Navicrawler is a Firefox add-on designed and developed by Mathieu Jacomy. The interface is located on the left of the currently browsed page.

Navicrawler works essentially by scraping the out-links of the visited websites (listed and stored as ‘Next Sites’). The researcher can then incorporate each website into the corpus, where it becomes an ‘in site’, or can reject it, and then it becomes an ‘out site’. The researcher can also describe the websites by adding tags.

The logic of exploration induced by Navicrawler combines browsing and crawling. Unlike automatic crawling, it allows the researcher to perceive the *context of links* and thus to avoid a blackbox effect. At the end of this exploration stage, s/he is able to export his/her corpus as a graph in which the nodes represent the websites and the edges stand for the links between them.

Step 2: Data enrichment (digital toolbox)

The social scientist plays a central role throughout the process of corpus building and description/enrichment. However, s/he can be assisted in the content analysis by automatic tools. Our research team developed a *digital toolbox* that makes possible various ‘enrichment processes’, which include:

– Retrieving from a list of URLs the information provided by the registrar about the registrant (owner of the domain name), especially his/her geographical location, about the server hosting the website, etc.
– Text-mining used on the index of the corpus in order to retrieve *named entities*: people, organizations, places, etc. (using Open Calais API).
– Recognition of the *language*s used in each website (and the distribution of languages in order to study multilingualism, an important issue in migration studies).

Step 3: Network visualization

In order to visualize the exploration data, in other words to map the corpus previously built, we use a graph-visualization software called *Gephi*, a project initiated and hosted at first by our research team. The software was developed by Mathieu Bastian, Mathieu Jacomy and Sébastian Heymann. This tool allows the user to spatialize and manipulate the corpus network. Two types of visualization are available:

– a spatialization based on the physical principle of attraction/repulsion (according to the presence or absence of a link between two nodes);
– a geographical spatialization that uses geocoded data: location of website owner, website addresses, servers, etc. (especially information retrieved during the data-enrichment stage).

Note: It must be stressed that the graph is a tool for the researcher, and not a 'photograph' of a given diaspora, while recognizing that our cartography represents a fragment of the Web of diasporas, a snapshot at a particular point in time.[5]

Step 4: Collaborative platform http://maps.e-diasporas.fr

The *maps.e-diasporas.fr* platform is a *collaborative* platform initially developed and implemented by Mathieu Jacomy and the ICT Migrations team for hosting the *e-Diasporas Atlas*. It is a tool for publishing and sharing research findings among scientific communities. The platform comprises *chapters* (in our case, the various diasporas) and provides for each of them the following data:

– *Maps*: browsable graphs of the corpus, with different views according to the fields of classification.
– *Raw data*: the empirical data (texts, videos, images, interviews, etc.) produced/retrieved and used during the research. The *e-Diasporas Atlas* is part of the more general 'digital humanities' project to provide access; it diffuses not only the research results but also the research data.
– *Statistics*: these are automatically generated from both the classification and the graph structure; they provide quantitative data about the relations between categories/actors. Statistics help strengthen the hypothesis formulated from the graph visualization.

Some significant findings

The *e-Diasporas Atlas* proposes at least two interpretation keys: (1) a topological key, centred on analysis of the connectivity between the actors on the Web; and (2) a quanti-qualitative key, which provides information derived from exploration of the contents of every site and from confrontation with the fieldwork and the expertise of all researchers involved in this project.

The comparison between geographical networks and networks observed on the Web seemed an obvious approach and gave rise to a few recurring observations, which included the following. The Web of diasporas does not match the geographical and demographical distribution of the dispersed populations. A large majority of e-diaspora sites are geolocated in North America and especially in the United States. This predominance is even more surprising when the presence on the Web strongly contrasts with the presence in the geographical areas, as was observed for the Palestinian, Nepalese or Egyptian diasporas. The online *hindutva* diaspora is thus very territorialized and symbolically linked to India, while operating from the United States. In this reallocation of geographical distribution due to presence or activity on the Web, the dominance of the English language and the majority practice of the Web favours English-speaking countries in the Web of diasporas. When English isn't dominant, the linguistic composition of the diaspora web reflects, above all, the diaspora's establishment in the host countries; very few websites operate exclusively in the language of origin.

Wishing to understand the collective life of the e-diasporas, we sought to identify clusters of expatriates on the Web and the exceptions to these clusters, their centrality, their hierarchies, their relations and their *assemblage*. The categorization of websites has enabled the identification of actors and clusters, and provides a glimpse, before going into further details, of: associations unaware of one another, bloggers creating their own world, activist groups or individuals seizing power on the Web and sometimes, as in the case of the Arab Spring, even managing to spark popular dissent and to impact on political events.

The absence of links between associations' websites cropped up repeatedly in the Web of diasporas. We also discovered, for instance, that relationships between different Lebanese organizations was problematic. The internal fragmentation of the Lebanese community sector appears to have been accentuated by the Web, and some alliances which exist on the ground are not visible in the graphs, as Houda Asal shows in this issue. Marta Severo observes that Egyptian associations based in different countries don't mention each other and have no common 'neighbour' websites. The websites of Palestinian associations are more frequently linked to frontier websites than amongst themselves.

When considering the categorization by publisher, the media are unquestionably the bridge nodes linking peripheral websites and institutional clusters. This may be explained by the fact that most associations' websites have a 'news' page related to the country of origin that links to articles published in major international or national media.

Mapping the e-diasporas enables us to analyse, among other things, the relationships undertaken and maintained by various diaspora actors with their homeland and with their institutions: clearly labelled links with a strong state, as in the case of France or India (an emergent state where we find a high visibility of government sites, which seek to attract the most privileged and influential migrants); one-way or even non-existent links, as in the case of Macedonia, Nepal and Lebanon for instance.

Two contributions to this special issue are particularly relevant for an understanding of the importance of neighbouring sites, and their relationship to diasporic websites.

In the corpus of the Palestinian diaspora put together by Anat Ben-David, 72% of the archived websites are neighbouring sites and 22% are diasporic websites. Together they form a densely knit network organized around two centres of gravity: the Palestinian cause and the Palestinian Territories. While analysing the dynamics of the Palestinian diaspora as it emerges on the Web, Ben-David observes that it is no longer defined around Palestine as a *place of origin*, but rather that it is constructed around Palestine as a *place of reference*. Ben-David argues that its organization is built less on a network of family, social and transactional ties between *communities* of Palestinians who have been dispersed to many places in the world, and more on global *advocacy networks* that transcend their immediate social networks. Its members are no longer only *Palestinians abroad*, but also *natives of the host countries* who identify with the Palestinian cause.

Concerning the Hindu diaspora, Ingrid Therwath identifies a corpus of websites which 'mirrors' the mother-organization as well as three types of frontier websites: American associations located in the institutional neighbourhood of the Sangh Parivar in the United States; generalist conservative American websites like Fox News or neo-liberal think tanks, located in the neighbourhood of blogs and non-institutional *hindutva* websites in India; and, between these two locations, Therwarth discovers a cluster of particularly virulent Jewish diaspora groups opposed to Muslims. Beyond the common Islamophobic discourse, this neighbourhood, which juxtaposes pro-*hindutva* groups and extremist Jewish groups, is particularly interesting in that it puts into contact diasporic groups from different regions.

The self-organization typical of Web networks facilitates the emergence of decentralized communities and acts as an ideal platform for different forms of mobilization. This is the case, for instance, of the Egyptian political e-diaspora, the nationalist religious *hindutva* movement, the memorial activism of French colonial repatriates, the Tunisian cyber-dissidence, the transnational solidarity mobilization in support of Tamil rights, or the boycott movement against Israeli commercial and cultural products.

In all these cases, we have observed that only part of a diaspora is active on the Web, and that among them only a small minority are involved in political action – despite their visibility and their dominance on the Web. The expansion of this activism depends on the events at hand and is generally associated with a specific context such as historical commemorations, radical regime changes or highly contested elections. In such cases, the movements are often instrumentalized by the home country, and gain access to the global public sphere through alternative media and non-diasporic actors.

Khachig Tölölyan wrote, in a famous article, 'Rethinking diaspora(s): stateless power in the transnational moment' (1996), that: '*Where once were dispersions, there now is diaspora*'. To conclude, we can paraphrase him by saying that *where once were diasporas, there now is the Web*… Most certainly, populations from different diasporas are now disseminated both throughout the world and on the Web – a phenomenon which deserves to be thoroughly studied.

Funding

In 2008 the *e-Diasporas Atlas* received funding from the Agence Nationale de la Recherche (STIC Contents), the Telecom ParisTech engineering school and the Fondation Maison des Sciences de l'Homme, Paris.

Notes

[1] Special thanks go to Matthieu Renault and Anne Rocha-Perazzo for their contribution.

[2] New communication and organization practices have produced a vast, moving e-corpus, whose exploration, analysis and archiving have never before been attempted. The *e-Diasporas*

Atlas is the first of its kind, and is the fruit of the efforts of more than 80 researchers worldwide, with some 8000 migrant websites archived and observed in their interactions.

The *e-Diasporas Atlas* was incubated and developed at the Fondation Maison des Sciences de l'Homme ICT Migrations program. Initiated and coordinated by Dana Diminescu, the project introduced digital methods into research on diasporas. This was made possible by the R&D innovations of Mathieu Jacomy and the technical coordination and training provided by Matthieu Renault. Some eighty researchers from diverse disciplines, laboratories and countries took part in the project. Several partners also contributed to its success: the Institut National de l'Audiovisuel, the Centre National de la Recherche Scientifique through its Migrinter laboratory, the Institut Mines-Telecom, Linkfluence and the design studio Incandescence. The *e-Diasporas Atlas* will continue to grow in the years to come.

[3] For general definitions of the technical terms used visit the subsection 'Learn more about our concepts, tools and methodology' at the website: http://www.e-diasporas.fr

[4] Diminescu D (2008) The connected migrant: An epistemological manifesto. *Social Science Information* 47(4): 565–579.

[5] The coloured graphs and subgraphs produced by and for the contributors to this Special Issue are not reproduced within the body of each article, but have been brought together in an appendix section located at the end of the issue and can be accessed at: http://dx.doi.org/10.1177/0539018412456918.

Author biography

Dana Diminescu, a practising sociologist, is an associate professor at TelecomParisTech. Her empirical work has enabled her to approach a variety of fields: e.g. uses of mobile telephone and voice IT, the Internet (tailing, archiving, mapping of the Web), and identifying digitalization technologies and m-transactions by migrants. Since 2003 she has been the scientific director of the research program ICT Migrations at the Fondation Maison des Sciences de l'Homme, Paris. This program, which she launched 10 years ago, has made major contributions to the theorization and analysis of the 'connected migrant'. She is also coordinator of the *e-Diasporas Atlas*, for which she developed a digital methodological chain and tools for mapping and analysing the occupation of the Web by diasporas.

Rethinking migration in the digital age: transglocalization and the Somali diaspora

SASKIA KOK[*] AND RICHARD ROGERS[†]

University of Amsterdam, Turfdraagsterpad 9,
1012 XT Amsterdam, The Netherlands
[*]S.L.E.Kok@uva.nl
[†]R.A.Rogers@uva.nl

Abstract *In this study, we examine the transnational networks of the Somali diaspora online. We explore the claims that the web signifies a shift towards a de-territorialized, transnational diaspora, which constructs its identity and engagement around a transnational imagined community. Based on a network and web content analysis, we assert that the claims about the transnational as the territorial locus of identity and engagement should be revisited. The analysis shows that the Somali diaspora's engagement has a specific multi-territorial topology through which information and resources are exchanged and a hybrid identity is constructed. Somalis' online engagement, however, is mainly directed towards community-based practices and social integration in their host-land, as opposed to transnational advocacy for the homeland. We argue that web data show a particular territorial arrangement and engagement, which we conceptualize as transglocalization, meaning local, networked formations existing alongside the national and transnational, each operating with awareness of the other yet acting separately. The study demonstrates that online network analysis offers promising approaches to diasporic social integration, policy-making and issue advocacy.*

Keywords NETWORK ANALYSIS, TRANSNATIONAL MIGRANTS, DIASPORA, DIGITAL METHODS

During the past twenty years, forced migration has accelerated substantively, generating a proliferation of diasporas pursuing economic betterment, escaping conflict and persecution and seeking human security (Van Hear 2003). One of the largest diasporas in the world is Somali, estimated at 1.2 million and living in countries such as the United Arab Emirates, United Kingdom, United States, Canada and Kenya (Hammond et al. 2011). While accurate data are hard to come by, it is estimated that the diaspora sends a staggering $1.6 billion back to Somalia annually, making it the fourth most remittance-dependent country worldwide (Kurz 2012). Empirical research surrounding the Somali diaspora has burgeoned, focusing not only on the large-scale remittance culture but also on issues such as host-land integration, homeland reconciliation, gender issues, clan culture and Islam (Collet 2007; Fangen 2006; Hopkins 2006). It has been found that Somalis residing in Western countries face difficulties of social integration and are often marginalized, being under- or unemployed and unable to utilize their education or job-related qualifications (Kleist 2008a). The presence of a discursive construction whereby Somalis are seen as 'very difficult to integrate', amplifies the ambiguous perception of Somalis, which partly results from 'derogatory ideas about Islam, female subordination and problematic masculinity' (Kleist 2008b: 307–23).

Indeed the Somali 'community' faces several difficulties when migrating abroad, one of which is that they do not form a single group defined by their predicaments (Harris 2004). Where the term 'community' implies cohesion or uniformity – which is certainly true for particular aspects such as their common language or Islam religion – their places of origin, prevalence of clan-based fragmentation, and individual experiences of war, diminish homogeneous conceptualizations, and complicate asylum claims. Therefore, community organizations often focus on dealing with struggles for political recognition and voicing a broader Somali interest. The latter is in keeping with the 'recognition turn' (Taylor 1994), where legitimacy is attributed (when successful) to social struggles driven by experiences of misrecognition or marginalization (Kleist 2008a). As such, it is argued that the identity of the so-called 'Somali diaspora' constitutes two different repertoires: on the one hand, Somalis are caught up in the suffering and marginalization of their life in exile; on the other, they are entangled in the invocations of a transnationally committed community dedicated to the development of the homeland (Kleist 2008a).

The networked, transnational dimensions of contemporary migration

During the 1990s, a theoretical approach emerged that conceptualized migration away from the homeland and host-land dichotomy, in favour of the perspective of transnationalism (Glick Schiller et al. 1995), which (Basch et al. 1994: 7) defined as the 'process by which migrants forge and sustain multi-stranded social relations that link together their countries of origin and settlement'. Hence, scholars recognized that migration is a dynamic process spanning various national and global geographies, cultures and political networks, as well as opportunity structures. From this perspective, members of diasporas are described as transmigrants (Faist 2000; Glick Schiller et al. 1995), and their multi-territorial behaviour as glocal or translocal (Friedman 2005; Giulianotti and Robertson 2007; Portes 2001). Recent scholarship has called for nuancing the relationship between the homeland, the host-land and the transnational. An example would be how sustained integration activities by diaspora groups may reinforce (positively or negatively), co-exist with, or substitute for, a transnational advocacy outlook (Erdal and Oeppen 2013; Hammond 2013; Tsuda 2012). In this study, we take up this invitation by turning to the space often analysed when seeking the transnational, that is, the web and new media technologies as well as the so-called connected migrant's use of them. Ultimately we find through an online analysis that the transnational, rather than being the de-territorial, interconnected state in which the Somali diaspora most meaningfully operates, is only one cluster in a larger network that is at the same time (and more significantly) local and national. These observations indicate that transnational engagement neither reinforces nor detracts from integration activities in national and local territories, and would be contiguous with research proclaiming that diasporic activity at the level of homeland, host-land and transnational, may co-occur (Tsuda 2012). However, our analysis has found a specific topology of multi-territorial engagement that is captured by the term transglocalization.

Indeed, intertwined with these spatial and territorial complexities of diaspora behaviour are the online activities made available by the web and new media technology (including social media platforms popularized by mobile apps). This allows for the emergence of a new diasporic space transcending territorial constraints. Several scholars have subsequently put forward the assertion that novel types of diasporic connectivity and a distinct online migrant community have emerged, referred to as digital, virtual or e-diaspora (Axel 2004; Brinkerhoff 2009; Diminescu 2008; Swaby 2013). The web and new media technology are said to augment a transnationalism in which diaspora groups are able to mobilize around a common diasporic identity, express their identity publicly and negotiate the terms of it, provide solidarity and material benefits to its members and engage in transnational political, economic and socio-cultural activities (Adamson 2012; Al-Sharmani 2006). Furthermore, the notion of *connected* migration points to networks organizing and circulating diasporic belonging activities (Diminescu 2008), whereby the emergence of the connected migrant (through mobility and connectivity) raises important questions for the study of the diaspora – particularly regarding relationships of identity, territorial or geographical belonging and (global) political engagement. Could diasporic territorial engagement, belonging and identity be constituted through an imagined transnational community? Which methods and techniques are suited for understanding connected migrant dynamics and the extent to which we are witnessing a shift away from the host-land–homeland dichotomy towards a de-territorialized, transnational diaspora? How might findings concerning connected migrants and their relationships between the local, national and transnational, inform policy-making both in terms of advocacy as well as homeland and host-land politics? We take these questions in turn, suggesting that the web and new media technologies are not only productive spaces for analysing claims about the transnational diaspora, including their advocacy, calling and identity formation activities, but are also effective in making data available for their study. These data, from websites, blogs and Facebook pages of self-identifying Somali groups referencing the diaspora or similar terms, are seen as indicative of connected migrants' diasporic activities, as we detail. We do not claim that these connected migrants represent the entire Somali diaspora or even the overall online Somali diaspora, which to our knowledge has yet to be sufficiently demarcated. Having said that, we have made a significant effort in that direction by collecting robust data about pertinent websites and Facebook pages from leading geographical places where the Somali diaspora are based, as we describe below.

We examine the transnational political engagement of the Somali diaspora, and its transnational linkages and global diaspora politics in particular, defined as the forms of transnational political engagement that are structured around a particularistic identity category (such as a national, ethnic, religious or sectarian identity) and a real or mythical 'homeland' (Adamson 2012). We further inquire into the extent to which the web may help aid not only transnational formations and engagement, but also, especially, their study. We employ a cross-disciplinary approach, drawing on literature from international relations and media studies. While the dynamic state of diaspora engagement has often been addressed in the literature as the 'triadic relationship', seeking to capture the ongoing dialogue between the 'home state', 'host-state' and the 'diaspora' (Vertovec 1997), this type of reasoning is further complicated by the emergence of a spatial understanding made available with web data. Ultimately, we propose to reconceptualize the way in which global diaspora politics are reconfigured on the web. We argue that the advent of the web and web data introduce new understandings of spatial configurations of the diaspora and its activities, which we dub transglocalization. Transglocalization is defined as the dynamic process of migration, traceable online, in which national networked formations exist alongside the local and the transnational, operating with knowledge and awareness of the other, yet acting separately. The process of transglocalization aids in the construction of diasporic identity for the connected migrant, where identity and interconnectedness refer to the multiplication of locations at all geographical levels, and where new networks of belonging and calling activities are available and sustained by new media technology (Diminescu 2008).

From a methodological perspective, we argue that this shift in diasporic activity indicates that the web may be seen as a distinct site of diasporic behaviour as well as a source of data. Somali diaspora groups maintain cultural bonds through websites and social media, reaching out to other Somali communities, locally, nationally and/or transnationally, providing a source of (homeland, host-land and transnational) calling activities that subsequently may be captured and studied. From an empirical perspective we argue that the web alters the space of global diaspora politics and identity production in that it furnishes not only knowledge and awareness of local, national and transnational diasporic formations (and thus the transglocal as defined above), but also what engages the members of the diaspora.

Following an examination of the existing literature on diasporas and transnationalism in relation to new media technology, we argue that previous studies have underplayed the empirical examination of online networks for the study of global diaspora politics and the Somali diaspora in particular. We subsequently employ *digital methods* for corpus building and analysis in a three-step process: web data capture, data enrichment, and network analysis and visualization (Rogers 2013). The empirical section shows two Somali diaspora corpora gathered from Google web search and Facebook, and an online network and web-content analysis, which are our means of examining the space and diasporic engagement. For each of the corpora, its key features are presented in relation to directionality of calling (homeland, host-land, transnational), geographical location, level of institutionalization, activity and cause. In the final section, we present the implications for future research and policy-making, whereby we demonstrate how digital methods provide approaches to the study of connected migrants, and how the outcomes may be employed for social integration, policy-making and issue advocacy, taking into account transnational practices and linkages to the host-land as well as homeland.

Transnational diasporic behaviour

During the 1970s, the term diaspora was increasingly used to denote almost everyone living away from his or her ancestral homeland. In his seminal work, John Armstrong (1976: 393) defines diaspora as 'any collectivity, which lacks a territorial base within a given polity'. More recently, other scholars have emphasized their topological characteristics (Cohen 1997) or the intrinsic transnational features and conceptions of diasporic collectivity (Adamson and Demetrious 2007; Anderson 1983), and often broadly define diasporas as the 'spatial dispersal of a people from an existing or imaginary homeland, maintaining a sense of collectivity over an extended period of time' (Bauman 2000; Kleist 2008a: 1129; Van Hear 1998).

As a result, the term diaspora shifts the focus towards 'issues of displacement and diasporic consciousness as well as to ideas of homeland and return' (Kleist 2008a: 1129). For example, Cohen (1997) argues that a diaspora is signified as a collective trauma, a banishment, where one dreams of home but simultaneously lives in exile. Diasporas are also viewed as an ethnic collectivity that not only has succeeded over time to sustain a common national, cultural or religious identity by maintaining a sense of internal cohesion and ties with the real or imagined homeland but also displays the ability to address collective interests of its members by establishing an organizational and transnational diaspora network (Adamson and Demetrious 2007).

The geographical dispersions of diasporas as a defining characteristic of migration flows has led to a cross-breeding of transnational and diaspora studies (Adamson 2005, 2012; Guarnizo et al. 2003; Kissau and Hunger 2010), influenced in part by the work of Keck and Sikkink and particularly their notion of transnational advocacy networks (TANs). While not describing diasporas *per se*, Keck and Sikkink (1999: 89) emphasize the global political importance of TANs, defined as networks that not only 'include those actors working internationally on an issue, who are bound together by shared values, and dense exchanges of information and services', but also use methods to influence policy-making distinct from those of other political entities. These transnational networks are, among other things, built on the sharing of information or what they call information politics.

Scholars have also studied the transnational engagement and cross-border political relationships of diasporas, including Adamson (2012) who views them as 'products or outcomes' of transnational mobilization. She argues that diasporas are political entrepreneurs who actively aim to construct and reify a transnational imagined community, referring in particular to Benedict Anderson's (1983) *Imagined communities*. With the advent of new global communication infrastructures these collective conceptions of nationality are utilized by the diaspora to construct political practices and identities, calling upon their transnational networks to create a de-territorialized social, cultural and political community, denoting the term 'diaspora' as prescriptive rather than descriptive.

Indeed, as Koinova notes, 'diasporas in the global age differ from nations of the modern age because they have multiple national identities and loyalties and are interlinked across the global' (Koinova 2010: 150). Just as the nation is not 'homogenous', dual-citizenship and other multiple loyalties are not yet conceptually integrated into the term diaspora (Koinova 2010). With this style of reasoning, Koinova's work is contiguous with other scholars who have emphasized that the study of transnationalism and governance should overcome the bias of so-called 'methodological nationalism' (Gonzales 2013; Wimmer and Glick Schiller 2002), and that scholars of international migration should reconsider the epistemological value of concepts such as nationalism, transnationalism or globalization (Vertovec 1999). Intertwined with these spatial and territorial complexities of diasporic behaviour are the online activities made available by web and new media technology, which allow for the emergence of a new diasporic space transcending territorial constraints.

The online dynamics of migration flows

New media technology has become a significant tool for creating connections among diaspora networks, civil society actors and policymakers (Newland 2010). These connections have enabled a different type of spatial connectivity and fostered the emergence of a distinct online migrant community (Brinkerhoff 2009; Diminescu 2008; Swaby 2013). As a result, a growing body of scholarship has explored the relationship between diasporas, technology and transnational migration (Brinkerhoff 2004, 2005, 2006, 2009; Diminescu 2008; Everett 2009) and their geographical dispersions, differentiating between such notions as transnational online communities, virtual/online diasporas, and ethnic online public spheres (Kissau and Hunger 2010). Others focus on the formation of collective identities, the socio-psychological dimensions of memory and displacement, political engagement, as well as the relationship with conflict and development. For instance, research has shown that diasporas often use social media to build an online diasporic public sphere in support of integration, or to help fill in the social void experienced when migrating to a different country (Diminescu et al. 2010; Ridings and Gefen 2004). The significance of online networks as a 'safe space' for diasporic interaction is also recognized, where diasporas can negotiate their sense of self, express their hybrid identities, or demarcate what it means to be a member (Brinkerhoff 2009; Swaby 2013). Hence, the transnational formations of diasporas aided by the web have had important implications for diasporic behaviour.

Some scholars have examined the socio-psychological dimensions of diasporic activity online. Of significance is the web's capacity as memory bearer as well as its role as active construction space for a (renewed) nation-state and identity, where the latter relates specifically to experiences of trauma, shame, denial, erasure and displacement (Bernal 2013; Estévez 2009). It is argued that new media technology alters experiences of displacement by the presence of (online) nostalgia, which secures a sense of belonging and is capable of erasing, or shrinking, time and space (Estévez 2009). Here, nostalgia is understood as the longing for a home that no longer exists, or never existed in the first place (Boym 2001). As a case in point, through the study of an online Eritrean war memorial, Awate.com, Bernal (2013) shows that members of Eritrean diasporas act as transnational citizens who construct their own histories of warfare, death, mourning and the silence that emerged during Eritrean liberation, when their own nation-state failed to do so.

New media technology has also been studied in relation to the engagement and mobilization of cross-border and diasporic political activism (Brinkerhoff 2009, Drissel 2006), as well as conflict and civil war (Brinkerhoff 2006, 2008). As a case in point, Drissel (2008: 90) argues that the Tibetan Buddhist youth have employed new media technologies to construct global online networks that have strengthened contacts and facilitated cross-border political activism by framing the Chinese government as 'the enemy' with 'bloody hands'. Ding (2007: 648) has shown that digital diasporas use the same means to shape the structure of transnational communication between the Chinese diaspora and their mainland, as well as Beijing's efforts to construct a favourable national image of 'new China'.

While the studies discussed above have recognized the relationship between the transnational perspective and new media technologies, they arguably have not aimed to capture the spatial complexity of diasporic networked behaviour in their empirical analyses. One of the few studies that have filled this gap is the e-Diasporas Atlas project led by Dana Diminescu (Diminescu 2012). The project's researchers examined 27 diaspora groups, using online mapping methods to build corpora and study the geography and occupations of diasporas (Ben-David 2012; Kumar 2012; Mazzuchelli 2012). The project is pioneering in approaching the diaspora as a hybridization between 'e'-social space and on-the-ground communities (Dumitriu 2012: 232). Here, the diaspora is studied together with its digital environments enabling a more complex spatial understanding, and the 'e-diaspora' becomes an ontological dimension to the diaspora as well as a significant part of the migrant's existence (Dumitriu 2012).

To sum up, the literature discussed above has made strides in studying diaspora activity online, yet some notes can be added regarding the methods. Most of the studies of the diaspora online employ qualitative methods for the study of single websites. Studies of this sort thereby fail to capture the larger scope as well as the characteristics of diaspora networks. Studies of websites such as Awate.com (Bernal 2013), Somalinet.com, TibetBoard (Brinkerhoff 2006, 2012), as well as of a small set of MySpace pages and blogs (Drissel 2008) are illustrative of this methodological tendency. Like the e-Diaspora Atlas, the aim of this study is to expand the corpus and contribute methodologically and conceptually to the study of global diaspora politics online.

Mapping the Somali diaspora online with digital methods

Scholars have asserted that new media technology such as the web, aids researchers in the study of new or existing social and cultural conditions, as well as the study of migration, thereby making (the connected migrants') web data an appropriate source for empirical investigation (Alonso and Oiarzabal 2010; Rogers 2013). Indeed, there has been a substantive development in methods using software to capture and repurpose digital data ('traces'), including the web analysis software *IssueCrawler* as well as Facebook analysis software *Netvizz* used here (Rieder 2013; Rogers 2010). As mentioned earlier, previous research surrounding digital diasporas has often employed a qualitative analysis of single (diaspora) websites. In our approach, we seek to expand the scope by joining with diaspora scholars studying 'online structures' (Adamson and Kumar 2014; Kissau and Hunger 2010). The aim here is thus to capture the network dynamics of the Somali community by building a corpus of migrant websites (for example, community associations, NGOs, forums, blogs and similar) which are organized by and around the Somali diaspora. We employ an online-network and web-content analysis, drawing on the methodology set out by the e-Diasporas Atlas program as well as the techniques developed at the Digital Methods Initiative (University of Amsterdam).

Where to begin when locating the Somali diaspora (online)? One of the most challenging research problems, affecting Somalis and non-Somalis alike, is scope; there is a wide variety of estimates of the size of the Somali population in their countries of residence, and no estimates of the Somali diaspora online (that we know of). Several official reports of the estimated Somali population abroad were employed as starting points for our selection of countries to study: United Kingdom (250,000-strong Somali diaspora), United States (10,000), Norway (24,000), Denmark (16,500), Sweden (15,500), the Netherlands (13,000), Germany (10,000), Italy (10,000) and Canada (50,000) (Kurz 2012; NHS 2011). The UNDP (Hammond et al. 2011) has studied the Somali diaspora by their residence in the capital cities, namely Dubai, London, Minneapolis, Nairobi, Oslo and Toronto. This study's country selection (and findings about cities) are based on the combination of the previously mentioned reports, resulting in the following countries under study: Kenya, United States, United Kingdom, the Netherlands, Canada, Norway, Sweden and Denmark. Another empirical note concerns the employed timeframe for the Facebook network-analysis; it was set at January 2012 to June 2014, a 30-month period. The rationale is based on the consideration that most of the diaspora-related pages came into existence during this time, and it was when the Facebook 'like' button (a metric for activity measures) gained popularity. While the Facebook's like button was introduced in 2010, its activity features (like, share, comment, like a comment) have only enjoyed widespread use since the end of the 2011 (Richmond 2011).

Web corpus building, data enrichment and network analysis and visualization

First, local domain search engines (google.co.uk, google.nl, google.co.ke, google.se, google.dk, google.no, google.ca and google.com) were queried for [Somali diaspora], [Somali community], [Somali diaspora interest group] and similar. URLs of the Somali diaspora websites were extracted from the engine results and entered into the IssueCrawler, both per country as well as collectively in one mother crawl. In the procedure, the IssueCrawler crawled and captured the outlinks of inputted sites, performed co-link analysis and outputted cluster maps. The results suggested that Facebook was the largest node in the (overall) network, thereby prompting the study of that platform. Facebook was queried in its search box for all the organization or group names as well as more generically for [Somali diaspora], [Somali community], [Somali diaspora interest group] and similar, and the pages found were liked, thereby enabling access to the pages' data using the Netvizz tool.

A platform specific content analysis was performed, whereby we determined most-engaged-with content on Facebook according to the sums of likes, shares, comments and liked comments, both per country as well as overall. Next, a classification scheme was developed for analysing the pages according to directionality of calling (homeland, host-land, transnational), geographical location, level of institutionalization, activity and cause (Ben-David 2012). The two classified corpora were imported into Gephi, the interactive exploration and visualization platform, where ties between pages and most-engaged-with content were determined (Gephi 2011). An inter-liked page analysis was undertaken and a page network was built, both per country as well as overall. For the Facebook inter-liked page analysis the top pages were chosen, or those pages with at least 40 likes. The inter-liked page analysis is a means by which to determine the ties between the pages (Gephi 2011). The network analysis enabled findings concerning a configuration of clusters we eventually named *transglocalization*, which included local, national as well as transnational ones, each loosely interlinked thereby showing each cluster's passing connection to every other link and its activities.

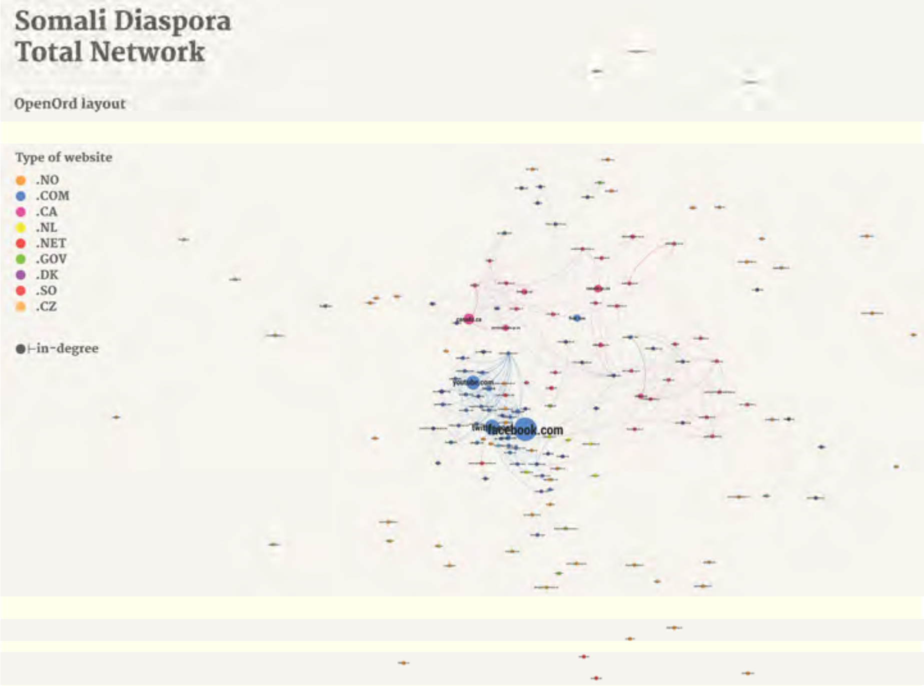
Findings from the web network and content analyses

We focus on two claims: that the web signifies a shift towards a de-territorialized, transnational diasporic engagement, and that diaspora construct their activities around a common transnational purpose including an imagined community. The former claim is examined through an online-network (hyperlink) analysis of websites and Facebook pages, and the latter through a web-content analysis.

To explore the claims, we first considered determinants of an online transnational network: interconnectivity of sites or pages and the qualities of the clusters (Diminescu 2008). The findings therefore first focus on the degree of interconnectivity transnationally, the relational aspects of the country clusters and directionality of calling. The directionality of calling is established through the categorization of the Somali corpora, and the description of activities listed on the diaspora websites that are directed towards host-land, homeland, host-land/homeland or considered transnational.

Figure 1 shows the complete hyperlink network of the Somali diaspora corpus crawled by the IssueCrawler. The network is comprised of a population of actors crawled from websites located in Kenya, Denmark, Canada, the Netherlands, Norway, UK and USA (The Swedish diaspora organizations did not have websites, but as seen below do have a select number of Facebook pages.) The overall network shows loose cohesion, with clusters of websites that are local, national as well as transnational, many of which are based in the USA and Canada.

Figure 1: Visual representation of the Somali web corpus and its interlinkings, consisting of 165 nodes and 336 edges



Issue Crawler data – rendered by Gephi. 26 June 2014.

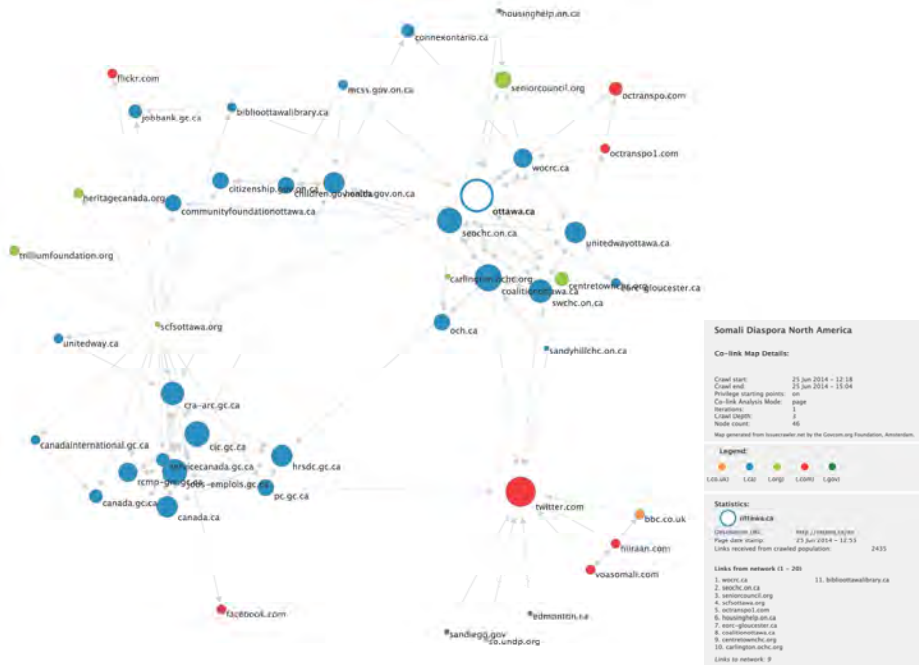
The US and Canadian based organizations link to Canadian governmental websites, citizen and immigration sites, and local community organizations such as the Coalition of Community Health and Resource Centres of Ottawa, indicating, at first glance, the prevalence of specific local and national geographical orientations, together with host-land integration activities. The network also points to Facebook as a central node. Figure 2 represents the Canadian and US networks, which point to social integration specifically (as opposed to a homeland calling orientation). The country networks of Kenya, the Netherlands, Norway, Denmark and the UK show little cohesion, indicating a lack of national ties between the Somali organizations.

The 165 Somali websites in the corpus were categorized according to directionality of calling. We again found the tendency for host-land calling activities (see Table 1), as 60 per cent of the organizations direct their activities towards the host-land, 26 per cent towards the homeland, 5 per cent to the host-land/homeland and 9 per cent are considered to be transnational.

Facebook analysis of the Somali diaspora with Netvizz

Since Facebook proved to be a significant node, a corpus of Somali diaspora pages was built. There were 61 Somali diaspora organizations present on Facebook, according to the sourcing technique described above (Table 2).

Figure 2: Visual representation of the Somali web corpus and its interlinkings for Canada and the USA



Issue Crawler data. 26 June 2014

Table 1: Classification of Somali websites in the web-corpus by directionality of calling, 25 June 2014 (N=165).

Directionality of calling	Homeland	Host land	Host land/ homeland	Transnational	Total
United States	4	30	0	2	36
The Netherlands	12	9	4	0	25
Norway	4	17	1	3	25
United Kingdom	8	11	3	1	23
Canada	0	15	0	2	17
Unknown/Transnational	7	5	0	3	15
Kenya	6	2	0	2	10
Denmark	0	8	1	0	9
Somalia	4	0	0	0	4
Ireland	0	1	0	0	1
Total	**44**	**98**	**9**	**13**	**165**
Percentage	**26**	**60**	**5**	**8**	**100**

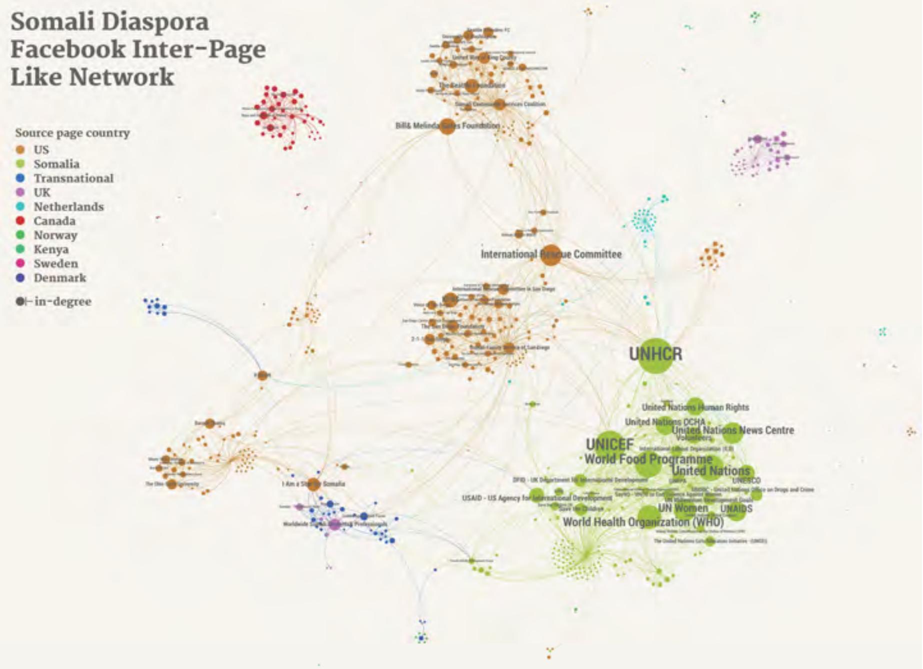
Note: The Somali diaspora websites located in Somalia and Ireland were found in the google.com and google.co.uk searches respectively.

Table 2: Summary statistics for the network analysis of Somali diaspora on Facebook, 615 nodes and 2021 connections

	Value
Average Path Length (degree of separation)	5.28
Average Degree (number of connections for each organization)	3.28
Clustering Coefficient (With 0 being no connections and 1 fully connected)	0.26
Graph Density (connections per org. compared with total organizations in network)	0.01
Network Diameter (longest graph distance between any two nodes)	13
Modularity (sub-network differentiation)	0.75
Number of Communities	53

An inter-liked page analysis was performed through Netvizz and visualized in Gephi (Figure 3), showing a rather loosely interconnected, yet well-clustered, network. In sum, the results suggest the lack of a transnational Somali network, owing to the low graph density, clustering co-efficient and connectivity levels overall. However, there are a fair number of communities from specific places.

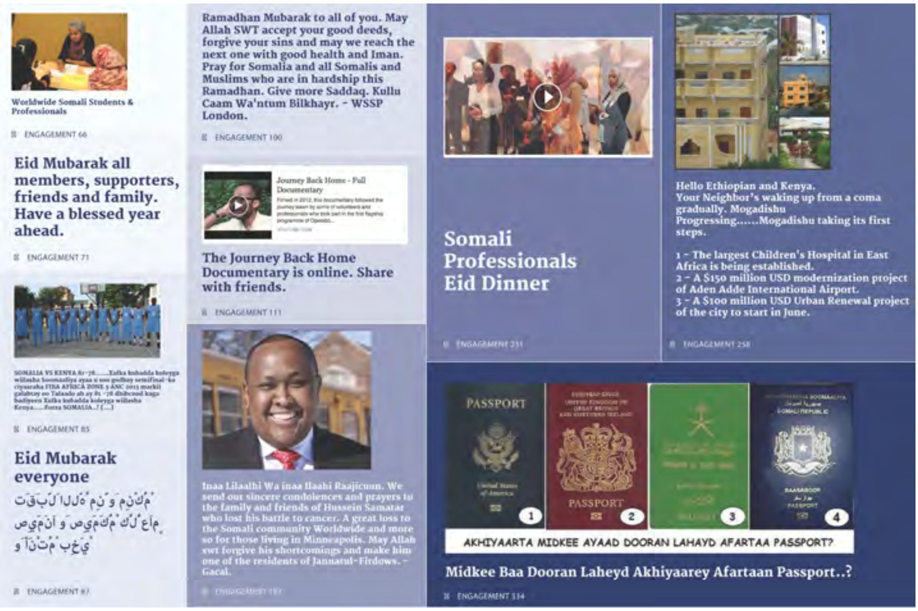
Figure 3: Network graph of the Somali Facebook corpus, where the total network (containing both diaspora and non-diaspora pages) consists of 615 nodes and 2021 edges.



Netvizz data – rendered by Gephi. 26 June 2014.

The analysis reveals a series of clusters, which could be interpreted as geographies. First, there is an 'integration geography' that is predominantly local, community-based, and not reaching out to international actors. The clusters are small US hubs such as San Diego, Seattle, Kansas City and Ohio (not Minneapolis, which is the transnational cluster's hub). There are further locally based clusters in Bristol (UK) as well as Ottawa (Canada). Second, there is a (smaller) cluster that revolves around 'I am a star for Somalia', the campaign organized by Somalis abroad to raise funds for the home country and to strengthen Somali identity. The organization (formally called American Refugee Committee, based in Minneapolis) links to Somali organizations located in the UK and the USA. Here Minneapolis becomes the launching point of the transnational rather than a local hub with community integration activities. This cluster links to transnational organizations, particularly umbrella NGOs, nonprofit organizations or large campaigns crossing borders. In fact, the 'I am a star' campaign organizes the only distinct transnational geography identified in this analysis, thereby complicating the claim that diaspora groups connect multiple points of dispersion (Adamson 2012; Adamson and Demetrious 2007; Kissau and Hunger 2010). The analysis of the most-engaged-with content on Facebook by each of these clusters bear out the findings described above. The favourite Canadian content was nearly exclusively concerned with integration activities, the US content with integration activities and transnational arts, and the Netherlands with host-land integration, homeland reconciliation and transnational activities, including the election victory by a Somali in a Dutch city council (see figures 4, 5 and 6).

Figure 4: Most-engaged-with content on Facebook by UK-based Somali Diaspora, January–June 2014



Worldwide Somali Students & Professionals

ENGAGEMENT 66

Eid Mubarak all members, supporters, friends and family. Have a blessed year ahead.

ENGAGEMENT 71

SOMALIA VS KENYA 81-78.......Xulka kubadda koleyga wiilasha Soomaaliya ayaa u soo gudbay semifinal-ka ciyaaraha FIBA AFRICA ZONE 5 ANC 2015 markii galabtay oo Talaado ah ay 81 -78 dhibcood kaga badiyeen Xulka kubadda koleyga wiilasha Kenya......Forza SOMALIA..! [...]

ENGAGEMENT 85

Eid Mubarak everyone
تَقَبَّلَ اللّٰهُ مِنّ وَ مِنْكُمْ
صِيَمْنَا وَ صِيَمَكُمْ كُلُّ عَامٍ
وَ أَنْتُمْ بِخَيْرٍ

ENGAGEMENT 87

Ramadhan Mubarak to all of you. May Allah SWT accept your good deeds, forgive your sins and may we reach the next one with good health and Iman. Pray for Somalia and all Somalis and Muslims who are in hardship this Ramadhan. Give more Saddaq. Kullu Caam Wa'ntum Bilkhayr. - WSSP London.

ENGAGEMENT 100

Journey Back Home - Full Documentary

The Journey Back Home Documentary is online. Share with friends.

ENGAGEMENT 111

Inaa Lilaalhi Wa inaa Ilaahi Raajicuun. We send our sincere condolences and prayers to the family and friends of Hussein Samatar who lost his battle to cancer. A great loss to the Somali community Worldwide and more so for those living in Minneapolis. May Allah swt forgive his shortcomings and make him one of the residents of Jannatul-Firdows. - Gacal.

ENGAGEMENT 187

Somali Professionals Eid Dinner

ENGAGEMENT 231

Hello Ethiopian and Kenya.
Your Neighbor's waking up from a coma gradually. Mogadishu Progressing......Mogadishu taking its first steps.

1 - The largest Children's Hospital in East Africa is being established.
2 - A $150 million USD modernization project of Aden Adde International Airport.
3 - A $100 million USD Urban Renewal project of the city to start in June.

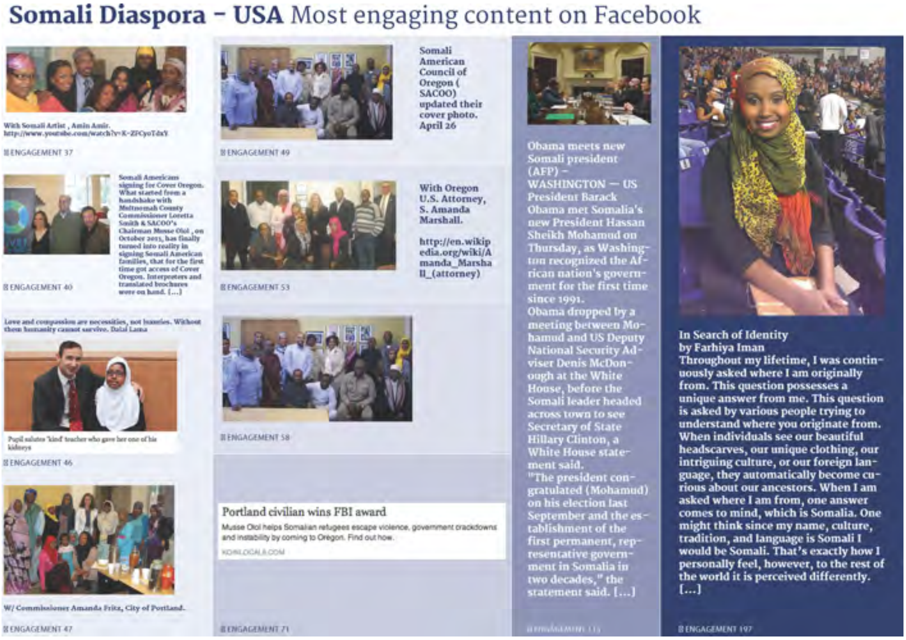
ENGAGEMENT 258

Midkee Baa Dooran Laheyd Akhiyaarey Afartaan Passport..?

ENGAGEMENT 334

Facebook data by Netvizz, analysed in Gephi and visualized as tree map. 26 June 2014.

Figure 5: Most-engaged-with content on Facebook by US-based Somali Diaspora, January–June 2014



Somali Diaspora - USA Most engaging content on Facebook

With Somali Artist , Amin Amir. http://www.youtube.com/watch?v=K-ZFCyoTdxY

ENGAGEMENT 37

Somali Americans signing for Cover Oregon. What started from a handshake with Multnomah County Commissioner Loretta Smith & SACOO's Chairman Musse Olol , on October 2013, has finally turned into reality in signing Somali American families, that for the first time got access of Cover Oregon. Interpreters and translated brochures were on hand. [...]

ENGAGEMENT 40

Love and compassion are necessities, not luxuries. Without them humanity cannot survive. Dalai Lama

Pupil salutes 'kind' teacher who gave her one of his kidneys

ENGAGEMENT 46

W/ Commissioner Amanda Fritz, City of Portland.

ENGAGEMENT 47

Somali American Council of Oregon (SACOO) updated their cover photo. April 26

ENGAGEMENT 49

With Oregon U.S. Attorney, S. Amanda Marshall.

http://en.wikipedia.org/wiki/Amanda_Marshall_(attorney)

ENGAGEMENT 53

ENGAGEMENT 58

Portland civilian wins FBI award

Musse Olol helps Somalian refugees escape violence, government crackdowns and instability by coming to Oregon. Find out how.

KOINLOCAL6.COM

ENGAGEMENT 71

Obama meets new Somali president (AFP) – WASHINGTON — US President Barack Obama met Somalia's new President Hassan Sheikh Mohamud on Thursday, as Washington recognized the African nation's government for the first time since 1991.
Obama dropped by a meeting between Mohamud and US Deputy National Security Adviser Denis McDonough at the White House, before the Somali leader headed across town to see Secretary of State Hillary Clinton, a White House statement said.
"The president congratulated (Mohamud) on his election last September and the establishment of the first permanent, representative government in Somalia in two decades," the statement said. [...]

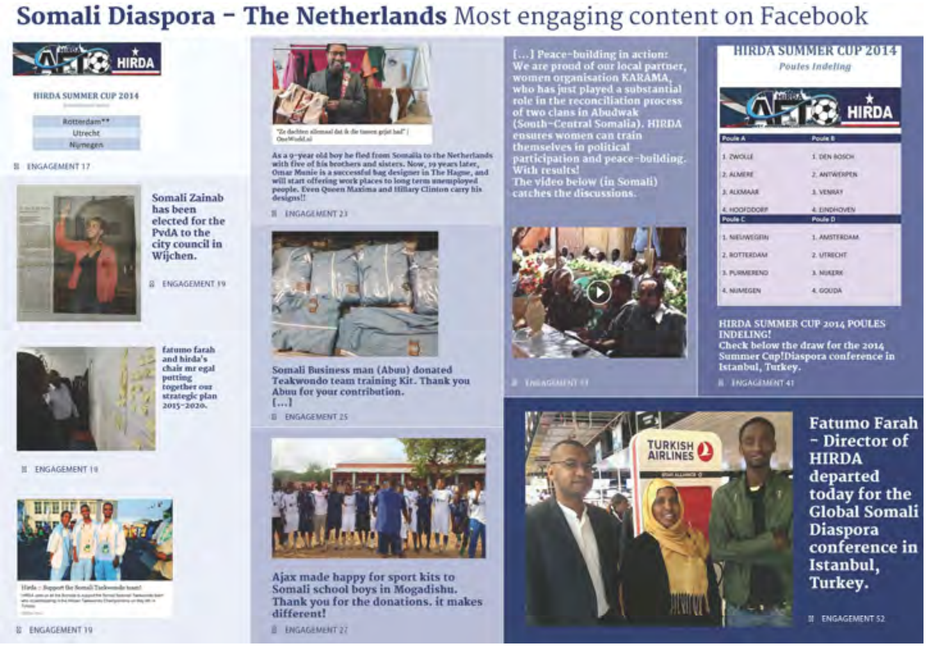
ENGAGEMENT 113

In Search of Identity
by Farhiya Iman
Throughout my lifetime, I was continuously asked where I am originally from. This question possesses a unique answer from me. This question is asked by various people trying to understand where you originate from. When individuals see our beautiful headscarves, our unique clothing, our intriguing culture, or our foreign language, they automatically become curious about our ancestors. When I am asked where I am from, one answer comes to mind, which is Somalia. One might think since my name, culture, tradition, and language is Somali I would be Somali. That's exactly how I personally feel, however, to the rest of the world it is perceived differently. [...]

ENGAGEMENT 197

Facebook data by Netvizz, analysed in Gephi and visualized as tree map. 26 June 2014.

Figure 6: Most-engaged-with content on Facebook by Dutch-based Somali Diaspora, January–June 2014



Somali Diaspora - The Netherlands Most engaging content on Facebook

HIRDA SUMMER CUP 2014

Rotterdam**
Utrecht
Nijmegen

ENGAGEMENT 17

Somali Zainab has been elected for the PvdA to the city council in Wijchen.

ENGAGEMENT 19

fatumo farah and hirda's chair mr egal putting together our strategic plan 2015-2020.

ENGAGEMENT 19

Hirda :: Support the Somali Taekwondo team!

ENGAGEMENT 19

"Ze dachten allemaal dat ik die tassen gejat had" | OneWorld.nl

As a 9-year old boy he fled from Somalia to the Netherlands with five of his brothers and sisters. Now, 19 years later, Omar Munie is a successful bag designer in The Hague, and will start offering work places to long term unemployed people. Even Queen Maxima and Hillary Clinton carry his designs!!

ENGAGEMENT 23

Somali Business man (Abuu) donated Teakwondo team training Kit. Thank you Abuu for your contribution. [...]

ENGAGEMENT 25

Ajax made happy for sport kits to Somali school boys in Mogadishu. Thank you for the donations. it makes different!

ENGAGEMENT 27

[...] Peace-building in action: We are proud of our local partner, women organisation KARAMA, who has just played a substantial role in the reconciliation process of two clans in Abudwak (South-Central Somalia). HIRDA ensures women can train themselves in political participation and peace-building. With results!
The video below (in Somali) catches the discussions.

ENGAGEMENT 33

HIRDA SUMMER CUP 2014

Poules Indeling

Poule A	Poule B
1. ZWOLLE	1. DEN BOSCH
2. ALMERE	2. ANTWERPEN
3. ALKMAAR	3. VENRAY
4. HOOFDDORP	4. EINDHOVEN
Poule C	**Poule D**
1. NIEUWEGEIN	1. AMSTERDAM
2. ROTTERDAM	2. UTRECHT
3. PURMEREND	3. NIJKERK
4. NIJMEGEN	4. GOUDA

HIRDA SUMMER CUP 2014 POULES INDELING!
Check below the draw for the 2014 Summer Cup!Diaspora conference in Istanbul, Turkey.

ENGAGEMENT 41

Fatumo Farah – Director of HIRDA departed today for the Global Somali Diaspora conference in Istanbul, Turkey.

ENGAGEMENT 52

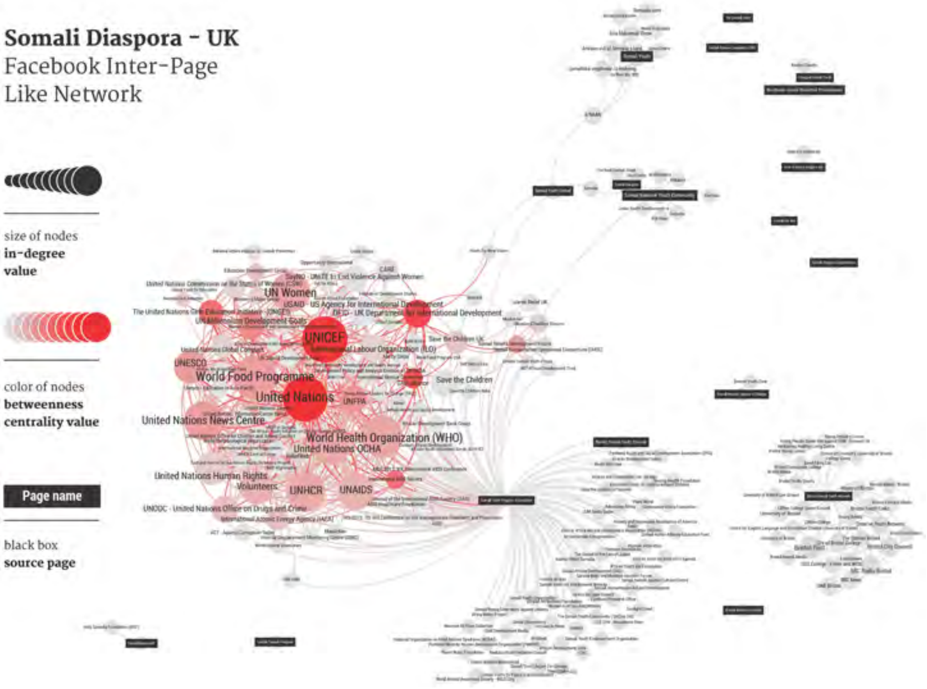
Facebook data by Netvizz, analysed in Gephi and visualized as tree map. 26 June 2014.

The individual country-based clusters reveal a series of findings with respect to cohesion or interconnectivity, and one could characterize each according to its specific array of host-land, homeland and/or transnational calling activities. The UK exhibits host-land activities and awareness of other national as well as international undertakings, having a distinctive local cluster centred on the Bristol Somali Youth Network, while linking to the musician K'Naan as well as Scandinavian Somali (youth) diasporic activities via a large international cluster (figure 7). Albeit larger, the US network is similar to the UK's in that there is a series of local hubs of host-land activities as well as a transnational cluster (out of Minneapolis) around 'I Am a Star'. The Canadian network also shows a cohesive and interconnected network, with a predominant emphasis on the local Ottawa community. The (small) country network of Sweden shows loose cohesion with hardly any interconnectivity, and has a directionality of calling primarily directed to the host-land. In Norway, three small clusters are visible, centred on the Somali Youth Community and a small human rights space, though most activities are again directed towards the host-land. Other country clusters have multiple directionalities of calling. The Dutch network is distinctive in that it is relatively cohesive with Hirda Netherlands (founded in 1996 by members of the Somali diaspora in the Netherlands to contribute to rehabilitation of the homeland), being the central node linking to host-land, homeland and international organizations (UN). The main organizations in the cluster are focused on social integration (Somali Jobs, Going Dutch), transnational Somali culture (such as K'naan, the Somali-Canadian musician, and the Somali Writers Collective) and UN organizations. The (small) network of Kenya is loosely connected but has a distinctive combination of actors, with directionalities of calling to both the homeland as well as host-land. Despite the complexity and variation of each country's diasporic geographical orientation and predominant directionalities of calling, it nevertheless may be concluded that the transnational, de-territorial activities are in the minority, generally, and may be located (so to speak) per country.

Web and Facebook content analysis findings

To explore the second claim concerning the extent to which the Somali diaspora construct its identity and engagement around a transnational imagined community, the data were further enriched by classifying the Somali organizations by type of organization, activity and cause. The total corpus (both corpora combined) is comprised of 226 organizations, consisting, for the most part, of community associations (52 per cent, 117 actors), non-profit/NGOs (30 per cent) and media (4 per cent) (Table 3). Organizational activity is centred on community-building (27 per cent) and youth (13 per cent) (Table 4). Furthermore, social integration (16 per cent, 36 actors), community-building (16 per cent) and women (5 per cent) are seen as significant organizational causes, demonstrating that the Somali diaspora is predominantly concerned with social and civic matters (Table 5). In contrast, the organizational type 'governmental institution' (1 per cent) and the activity 'politics' both rank relatively low (1 per cent), showing a lack of (formal) political themes among the Somali diaspora. When examining the corpora separately the findings are relatively similar, as most of the organizations are community associations and non-profit/NGOs focused on social integration, community networking and youth activities, with single-digit concern for causes like women's issues and human rights. This is not surprising given the widespread practice of female genital mutilation. Thus Somali diasporic networks differ from transnational issue networks in that they are more concerned with host-land integration activities than issue advocacy *per se* as their main mobilizing principle.

Figure 7: Network graph of the UK-based Somali diaspora on Facebook



Facebook data by Netvizz, analysed and visualized by Gephi. 26 June 2014.

Table 3: Classification of Somali diaspora Facebook and Web corpora by organization type (N=226)

Organization Type	Quantity	Percentage
Community Association	117	52.0
Non-profit/NGO	68	30.2
Unspecified	23	10.2
Media	10	4.4
Business	3	0.9
Government	2	0.9
Education	2	0.9
Industrial	1	0.5
Total	**226**	**100.0**

Table 4: Classification of Somali diaspora Facebook and web corpora by type of activity (N = 226)

Type of Activity	Quantity	Percentage
Community (building, networking)	61	26.9
Youth	30	13.3
Unspecified	23	10.2
Women	20	8.9
Human Rights	17	7.5
Education	10	4.4
Diaspora	9	4.0
Social integration	7	3.1
Media	7	3.1
Refugee	7	3.1
Children	4	1.8
State-building	4	1.8
Politics	3	1.3
Culture	3	1.3
Peace	3	1.3
Solidarity	3	1.3
Campaign	2	0.9
Agriculture/Sustainability	2	0.9
Sport	2	0.9
Umbrella	2	0.9
Business	2	0.9
Health	1	0.4
Law	1	0.4
Research	1	0.4
Social	1	0.4
Information	1	0.4
Total	**226**	**100.0**

Note: the classification results of the Somali diaspora corpora by type of activity and main cause (Table 5) are often times the same, but can also differ. For example, an organization focusing on women as their main type of activity could target social integration as their main cause.

Table 5: Classification of Somali diaspora Facebook and web corpora by main cause (N=226)

Main Cause	Quantity	Percentage
Social Integration	36	15.9
Community building	35	15.5
Unspecified	25	11.1
Women	12	5.3
Education	12	5.3
News	11	4.9
Human rights	7	3.1
Awareness	7	3.1
Welfare	6	2.7
Youth	5	2.2
Support	5	2.2
State-building	5	2.2
Civic Engagement	5	2.2
Politics	5	2.2
Culture	4	1.8
Diaspora	4	1.8
Solidarity	4	1.8
Sport	4	1.8
Leadership	4	1.8
Agriculture Sustainability	3	1.3
Children	3	1.3
Campaign	3	1.3
Law	3	1.3
Peace	3	1.3
Family	3	1.3
Development	2	0.9
Poverty/suffering	2	0.9
Sharing grievances	2	0.9
Counselling	1	0.4
Health	1	0.4
Housing	1	0.4
Protection	1	0.4
Refugee	1	0.4
Research	1	0.4
Total	**226**	**100.0**

Does the web show a shift towards a de-territorialized, transnational diaspora for the purpose of influencing policy-making or reimagining a Somali community? From the network analyses of the Somali diaspora websites as well as Facebook pages, the prospect of the 'transnational' organizing the Somali diaspora may be specified further. First, the diaspora networks comprise mainly national and local clusters, with relatively loose connections between them. There is a separate transnational cluster, which is smaller than the national and local ones, but nevertheless operates, as described in the literature, as advocacy – mobilizing and inter-linking sets of actors with shared values and a (homeland oriented) cause. So transnational issue networks exist, but issues or causes are not the main organizing entity of the diaspora overall, indicating that both claims stated earlier should be further elaborated. Scholars have asserted that diasporas function as an outcome of transnational mobilization constructing and reifying a transnational community; or, where TANs are utilized, they function by using the power of information, ideas or strategies to alter the context in which states create policies. The empirical findings suggest a rather different, more community-based explanation. Furthermore, the empirical findings confirm the assertion that tracing the presence and activities of the connected migrant explains more specifically the idea of diasporic communities living in the hybridized geography between the local communities and e-social space, made available by new media (Diminescu 2008; Dumitriu 2012).

Despite the proliferation of research examining how the internet enables fertile ground for the engagement of political activism and cross-border political relationships (Brinkerhoff 2009; Drissel 2008), members of the Somali diaspora appear far more concerned with social integration activities. Whilst Somalis are often found in marginalized positions, being underemployed and unable to utilize their job qualifications (Kleist 2008a), the empirical findings point towards a proactive attitude whereby individual achievement and integration are particularly rewarded. The Somali diaspora's online engagement can thus be conceptualized in the context of community-building efforts, in which they aim to fill a social void, imbued with recognition on the one hand, and practicing another form of citizenship on the other, which has also been noted in the literature (Bernal 2013; Diminescu et al. 2010; Ridings and Gefen 2004).

Conclusion: transglocalization, connected migrants and constituency-building

As noted previously, the web supposedly encourages the globalization of diasporic activities, yet simultaneously aids, seemingly more so, in the establishment of local practices. Some theorists have defined this dual process as glocalization (Friedman 2005; Giulianotti and Robertson 2007) by asserting that local cultures critically adapt or resist global phenomena through the creation of community-based polities. In a similar fashion, researchers have asserted that engagement with the transnational, local and national, including sustained integration activities, may co-exist side-by-side as a simultaneous process (for example, Tsuda 2012). Indeed, it is perceived that, in transnational networks, multiple identities and loyalties are interlinked across the global (Koinova 2010) and both notions of locality and spatiality need to be viewed as essential prerequisites surrounding processes of migration (Glick Schiller and Çaglar 2008). Furthermore, the social form of diaspora engagement has often been addressed in the literature as the 'triadic relationship', seeking to capture the ongoing dialogue between the host-state, home state, and the dispersed diaspora in which the tensions of divided loyalties and political orientations, as well as economic strategies of transnational engagement, are considered to be significant (Vertovec 1997). While these conceptions of 'social form' and engagement explain the diversity of diasporic behaviour, this type of reasoning is further complicated by the emergence of new spatial understandings made available with web data.

To capture these novel diasporic formations, the concept *transglocalization* is introduced. This is defined as the dynamic state of migration, traceable online, in which national networked formations exist alongside the local as well as the transnational, each operating with knowledge and awareness of the other yet acting separately. Future research should therefore take into account that the study of contemporary diasporic behaviour on the web fosters more complex multi-territorial formations and relationships between place and belonging.

The concept of the connected migrant in a transglocal state also offers a new terrain for policy-makers in the home state, the host state and the transnational community. Proactive diaspora engagement has been recognized by several scholars and has accordingly been labelled the 'diaspora option', encompassing several potential approaches that refer specifically to skilled diasporas as an asset to be captured (Alonso and Oiarzabal 2010; Meyer et al. 1997). This policy framework centres specifically on the perspective of the sending and receiving state as well as that of the migrant, by focusing on such notions as remittance capture, diaspora networking and diaspora integration (Gamlen 2005). From this perspective, it is recognized that diasporas do not necessarily constitute a threat to the sovereign nation-state, but rather can be viewed as potential partners to local communities and nation-states alike in promoting the enforcement of national and international law, and enhancing good governance (Brinkerhoff 2005).

Research has also shown, however, that identity groups often lack formal (inter)-national representation such as membership in the UN or representation in government and, as a result, rely on their dispersed members for external support (Demmers 2007). Most immigrants have few opportunities to inform and improve the policies that affect them on a daily basis. The Migration Integration Policy Index (measuring Europe and North America) reports that 11 countries, mostly in Eastern Europe, still have laws denying immigrants basic political liberties (MIPEX 2010). The success and ability of diaspora advocacy and integration therefore depends heavily on the political system of the 'target country' (Newland 2010: 6), logic of access and the migrants' social ties or education. For example, our analysis found that the Netherlands was a country where members of the Somali diaspora have gained political office (the announcement of which on Facebook was among the most-engaged-with content).

The connected migrant, though, holds potential to contribute meaningfully to the community development in the host-land. Their focus of activities could also be seen to play an important national policy role in encouraging social integration, at least this was found empirically. Second, it was observed that the 'transnational' diaspora is more of a cluster than an interlinked global community, with tenuous ties to the homeland. This raises the question of how to strengthen the transnational advocacy work, and its link to the homeland. Transnational and homeland-tied practices could tap into the knowledge and network of diasporas to solicit information and cultural and technical expertise for the purpose of constituency building and coordination, to better reach other diaspora communities that bear similar skills and expertise (Alonso and Oiarzabal 2010). The connected migrant should therefore be seen as a potential political actor, bearing significant resources for host-land, transnational and, to a lesser extent, homeland politics. The digital footprints they leave behind may serve as a valuable source for policy-makers concerned with both the places of social integration, or the lack thereof, and transnational engagement, including the activities that resonate across multiple country-clusters. The contributions of the connected migrants should thus be seen as multi-layered, whereby it is essential to identify local and transnational practices and calling activities that drive processes of integration, recognition and issue advocacy as well as tie-making with the homeland.

Acknowledgements

We would like to express our appreciation to the attendants of the Digital Methods Summer School 2014 for their useful comments, and Gabriele Colombo for analysis and graphic design.

References

Adamson, F. B. (2005) ‘Globalisation, transnational political mobilisation, and networks of violence’, *Cambridge Review of International Affairs*, 18 (1), 31–49, doi: 10.1080/09557570500059548.

Adamson, F. B. (2012) ‘Constructing the diaspora: diaspora identity politics and transnational social movements’, in P. Mandaville and T. Lyons (eds) *Politics from afar: transnational diasporas and networks*, New York: Columbia University Press, 25–42.

Adamson, F. B. and M. Demetrious (2007) ‘Remapping the boundaries of “state” and “national identity”: incorporating diasporas into IR theorizing’, *European Journal of International Relations*, 13 (4), 489–526, doi: 10.1177/1354066107083145.

Adamson, F.B. and P. Kumar (2014) ‘Imagined communities 2.0: space and place in Tamil, Sikh and Palestinian online identity politics’, paper presented at the 55th annual meeting of the International Studies Association, Toronto, Canada, 26–29 March.

Al-Sharmani, M. (2006) ‘Living transnationally: Somali diasporic women in Cairo’, *International Migration*, 44 (1), 55–77, doi: 10.1111/j.1468-2435.2006.00355.x.

Alonso, A. and P. J. Oiarzabal (2010) *Diasporas in the new media age: identity, politics, and community*, Reno: University of Nevada Press.

Anderson. B. (1983) *Imagined communities: reflections on the origin and spread of nationalism*, London: Verso.

Armstrong, J. A. (1976) ‘Mobilized and proletarian diasporas’, *American Political Science Review*, 70 (2), 393–408, doi:10.2307/1959646.

Axel, B. K. (2004) ‘The context of diaspora’, *Cultural Anthropology*, 19 (1), 26–60, doi: 10.1525/can.2004.19.1.26.

Basch, L., N. Glick Schiller and C. Szanton Blanc (1994) ‘Towards a definition of transnationalism,’ in L. Basch, N. Glick Schiller and C. Szanton Blanc (eds) *Nations unbound: transnational projects, postcolonial predicaments and deterritorialized nation-states*, Amsterdam: Gordon and Breach, 1–24.

Baumann, M. (2000) ‘Diaspora: genealogies or semantics and transcultural comparison’, *Numen*, 47 (3), 313–37, doi: 10.1163/156852700511577.

Ben-David, A. (2012) ‘The Palestinian diaspora on the web: between de-territorialization and re-territorialization’, *e-Diasporas Atlas*, available at: http://docplayer.net/11138468-The-palestinian-diaspora-on-the-web-between-de-territorialization-and-re-territorialization.html.

Bernal, V. (2013) ‘Diaspora, digital media, and death counts: Eritreans and the politics of memorialisation’, *African Studies*, 72 (2), 246–64, doi: 10.1080/00020184.2013.812875.

Boym, S. (2001) *The future of nostalgia*, New York: Basic Books.

Brinkerhoff, J. M. (2004) ‘Digital diasporas and international development: Afghan-Americans and the reconstruction of Afghanistan’, *Public Administration and Development*, 24 (5), 397–413, doi: 10.1002/pad.326.

Brinkerhoff, J. M. (2005) ‘Digital diasporas and governance in semi-authoritarian states: the case of the Egyptian Copts’, *Public Administration and Development*, 25 (3), 193–204, doi: 10.1002/pad.364.

Brinkerhoff, J. M. (2006) ‘Digital diasporas and conflict prevention: the case of somalinet.com’, *Review of International Studies*, 32 (1), 25–47, doi: 10.1017/S0260210506006917.

Brinkerhoff, J. M. (2008) ‘Diaspora identity and the potential for violence: toward an identity-mobilization framework’, *Identity: An International Journal of Theory and Research*, 8 (1), 67–88, doi: 10.1080/15283480701787376.

Brinkerhoff, J. M. (2009) ‘Creating an enabling environment for diasporas’ participation in homeland development’, *International Migration*, 50 (1), 75–95, doi: 10.1111/j.1468-2435.2009.00542.x.

Brinkerhoff, J. M. (2012) ‘Digital diasporas’ challenge to traditional power: the case of TibetBoard’, *Review of International Studies*, 38 (1), 77–95, doi: 10.1017/S02602105100 01737.

Cohen, R. (1997) *Global diasporas: an introduction*, London: UCL Press and Seattle: University of Washington Press.

Collet, B. A. (2007) ‘Islam, national identity and public secondary education: perspectives from the Somali diaspora in Toronto, Canada’, *Race Ethnicity and Education*, 10 (2), 131–53, doi: 10.1080/13613320701330668.

Demmers, J. (2007) ‘New wars and diasporas: suggestions for research and policy’, *Journal of Peace, Conflict & Development*, 11, 1–26, available at: http://goo.gl/uieUnH.

Diminescu, D. (2008) ‘The connected migrant: an epistemological manifesto’, *Social Science Information*, 47 (4), 565–79, doi: 10.1177/0539018408096447.

Diminescu, D. (2012) ‘Introduction: digital methods for the exploration, analysis and mapping of e-diasporas’, *Social Science Information*, 51 (4), 451–8, doi: 10.1177/0539018412456918.

Diminescu, D., M. Renault, M. Jacomy and C. D’Iribarne (2010) ‘Le web matrimonial des migrants’, *Réseaux*, 1 (159), 15–56, doi: 10.3917/res.159.0015.

Ding, S. (2007) ‘Digital diaspora and national image building: a new perspective on Chinese diaspora study in the age of China’s rise’, *Pacific Affairs*, 80 (4), 627–48, doi: 10.5509/2007804627.

Drissel, D. (2008) ‘Digitizing Dharma: computer-mediated mobilizations of Tibetan Buddhist youth’, *The International Journal of Diversity in Organizations, Communities and Nations*, 8 (5), 79–92, available at: http://ijd.cgpublisher.com/product/pub.29/prod.742.

Dumitriu, D. L. (2012) ‘E-Diasporas Atlas: exploration and cartography of diasporas in the digital network’, *Romanian Journal of Communication and Public Relations*, 14 (S4), 231–4, available at: http://codipo.ro/dox/dumitriu_28.pdf.

Erdal, M. B. and C. Oeppen (2013) ‘Migrant balancing acts: understanding the interactions between integration and transnationalism’, *Journal of Ethnic and Migration Studies*, 39 (6), 867–84, doi: 10.1080/1369183X.2013.765647.

Estévez, S. M. (2009). ‘Is nostalgia becoming digital? Ecuadorian diaspora in the age of global capitalism’, *Social Identities*, 15 (3), 393–410, doi: 10.1080/13504630902899366.

Everett. A. (2009) *Digital diaspora: a race for cyberspace*, Albany: SUNY Press.

Faist, T. (2000) 'Transnationalization in international migration', *Ethnic and Racial Studies*, 23 (2), 189–222, doi: 10.1080/014198700329024.

Fangen, K. (2006) 'Humiliation experienced by Somali refugees in Norway', *Journal of Refugee Studies*, 19 (1), 69–93, doi: 10.1093/jrs/fej001.

Friedman. T. L. (2005) *The world is flat*, New York: Farrar, Straus and Giroux.

Gamlen A. (2005) 'The brain drain is dead; long live the New Zealand diaspora', working paper 05-10. Oxford: COMPAS, University of Oxford, available at: www.compas.ox.ac.uk/media/WP-2005-010-Gamlen_Diaspora_New_Zealand.pdf.

Gephi (2011) 'Modularity', available at: https://wiki.gephi.org/index.php/Modularity.

Giulianotti, R. and R. Robertson (2007) 'Forms of glocalization: globalization and the migration strategies of Scottish football fans in North America', *Sociology*, 41 (1), 133–152, doi: 10.1177/0038038507073044.

Glick Schiller, N. and A. Çaglar (2008) 'Migrant incorporation and city scale: towards a theory of locality in migration studies', Willy Brandt series of working papers in international migration and ethnic relations, Malmö: Malmö University, available at: http://hdl.handle.net/2043/5935.

Glick Schiller, N., L. Basch and C. Szanton Blanc (1995) 'From immigrant to transmigrant: theorizing transnational migration', *Anthropological Quarterly*, 68 (1), 48–63, doi: 10.2307/3317464.

Gonzales. T. (2013) 'Transnationalism', CoHaB (Diasporic Constructions of Home and Belonging), interdisciplinary research network project, available at: http://www.itn-cohab.eu/wiki/transnationalism.

Guarnizo. L. E., A. Portes and W. Haller (2003) 'Assimilation and transnationalism: determinants of transnational political action among contemporary migrants', *The American Journal of Sociology*, 108 (6), 1211–48, doi: 10.1086/375195.

Hammond, L. (2013) 'Somali transnational activism and integration in the UK: mutually supporting strategies', *Journal of Ethnic and Migration Studies*, 39 (6), 1001–17, doi: 10.1080/1369183X.2013.765666.

Hammond, L., M. Awad, A. I. Dagane, P. Hansen, C. Horst, K. Menkhaus and L. Obare (2011) 'Cash and compassion: the role of the Somali diaspora in relief, development and peace-building', UNDP report, available at: www.refworld.org/pdfid/4f61b12d2.pdf.

Harris, H. (2004) 'The Somali community in the UK: what we know and how we know it', report commissioned by the Information Centre about Asylum and Refugees in the UK (ICAR), available at: www.icar.org.uk/somalicommunityreport.pdf.

Hopkins, G. (2006) 'Somali community organizations in London and Toronto: collaboration and effectiveness', *Journal of Refugee Studies*, 19 (3), 361–80, doi: 10.1093/jrs/fel013.

Keck, M. E. and K. Sikkink (1999) 'Transnational advocacy networks in international and regional politics', *International Social Science Journal*, 51 (159), 89–101, doi: 10.1111/1468-2451.00179.

Kissau, K. and U. Hunger (2010) 'The internet as a means of studying transnationalism and diaspora', in R. Bauböck and T. Faist (eds) *Diaspora and transnationalism: concepts, theories and methodology*, Amsterdam: Amsterdam University Press, 245–66.

Kleist, N. (2008a) 'In the name of diaspora: between struggles for recognition and political aspirations', *Journal of Ethnic and Migration Studies*, 34 (7), 1127–43, doi: 10.1080/13691830802230448.

Kleist, N. (2008b) 'Mobilising "the diaspora": Somali transnational political engagement', *Journal of Ethnic and Migration Studies*, 34 (2), 307–23, doi: 10.1080/13691830701823855.

Koinova, M. (2010) 'Diasporas and international politics: utilising the universalistic creed of liberalism for particularistic and nationalist purposes', in R. Bauböck and T. Faist (eds) *Diaspora and transnationalism: concepts, theories and methods*, Amsterdam: Amsterdam University Press, 149–66.

Kumar, P. (2012) 'Sikh narratives: an analysis of virtual diaspora networks', *e-Diasporas Atlas*, available at: http://www.e-diasporas.fr/working-papers/Kumar-Sikhs-EN.pdf.

Kurz, R. W. (2012) 'Europe's Somali diaspora: both a vulnerability and a strength', Foreign Military Studies Office/EUCOM project report, available at: http://goo.gl/ySsw8r.

Mazzuchelli, F. (2012) 'What remains of Yugoslavia? From the geopolitical space of Yugoslavia to the virtual space of the web Yugosphere', *e-Diasporas Atlas*, available at: www.e-diasporas.fr/working-papers/Mazzucchelli-Yugosphere-EN.pdf.

Meyer, J.-B., J. Charum, D. Bernal, J. Gaillard, J. Granés, J. Leon, A. Montenegro, A. Morales, C. Murcia, N. Narvaez-Berthelemot, L. S. Parrado and B. Schlemmer (1997) 'Turning brain drain into brain gain: the Colombian experience of the diaspora option', *Science, Technology and Society*, 2 (2), 285–315, doi: 10.1177/097172189700200205.

MIPEX (Migration Integration Policy Index) (2010) 'Political participation', available at: www.mipex.eu/political-participation.

NHS (National Household Survey) (2011) 'National Household Survey profile', Statistics Canada catalogue, report 99-004-XWE, 11 September, available at: www12.statcan.gc.ca/nhs-enm/2011/dp-pd/prof/index.cfm.

Newland, K. (2010) 'Voice after exit: diaspora advocacy', Migration Policy Institute report, available at: www.migrationpolicy.org/sites/default/files/publications/diasporas-advocacy.pdf.

Portes, A. (2001) 'Introduction: the debates and significance of immigrant transnationalism', *Global Networks*, 1 (3), 181–93, doi: 10.1111/1471-0374.00012.

Richmond, R. (2011) 'As "like" buttons spread, so do Facebook's tentacles', *New York Times*, 27 September, available at: http://bits.blogs.nytimes.com/2011/09/27/as-like-buttons-spread-so-do-facebooks-tentacles/.

Ridings, M. R. and D. Gefen (2004) 'Virtual community attraction: why people hang out online', *Journal of Computer Mediated Communication*, 10 (1), doi: 10.1111/j.1083-6101.2004.tb00 229.x.

Rieder, B. (2013) 'Studying Facebook via data extraction: the Netvizz application', *WebScience*, proceedings of the 5th annual ACM Web Science Conference, 2–4 May, Paris, 346–55, doi: 10.1145/2464464.2464475.

Rogers, R. (2010) 'Mapping public web space with the Issuecrawler', in C. Brossard and B. Reber (eds) *Digital cognitive technologies: epistemology and knowledge society*, London: ISTE, 115–26.

Rogers, R. (2013) *Digital methods*, Cambridge, MA: MIT Press.

Swaby, N. (2013) 'Digital diaspora', CoHaB (Diasporic Constructions of Home and Belonging), interdisciplinary research network project, available at: www.itn-cohab.eu/wiki/digital-diaspora.

Taylor, C. (1994) 'The politics of recognition', in C. Taylor and A. Gutmann (eds) *Multiculturalism: examining the politics of recognition*, Princeton, NJ: Princeton University Press, 25–74.

Tsuda, T. (2012) 'Whatever happened to simultaneity? Transnational migration theory and dual engagement in sending and receiving countries', *Journal of Ethnic and Migration Studies*, 38 (4), 631–49, doi: 10.1080/1369183X.2012.659126.

Van Hear, N. (1998) *New diasporas: the mass exodus, dispersal and regrouping of migrant communities*, Seattle, Washington: University of Washington Press.

Van Hear, N. (2003) 'Refugee diasporas, remittances, development and conflict', Migration Policy Institute publication, 1 June, available at: www.migrationpolicy.org/article/refugee-diasporas-remittances-development-and-conflict.

Vertovec, S. (1997) 'Three meanings of "diaspora," exemplified among South Asian religions', *Diaspora: Journal of Transnational Studies*, 6 (3), 277–99, doi: 10.1353/dsp.1997.0010.

Vertovec, S. (1999) 'Conceiving and researching transnationalism', *Ethnic and Racial Studies*, 22 (2), 447–62, doi: 10.1080/014198799329558.

Wimmer, A., and N. Glick Schiller (2002) 'Methodological nationalism and beyond: nation-state building, migration and the social sciences', *Global Networks*, 2 (4), 301–34, doi: 10.1111/1471-0374.00043.

MIGRATION AND NEW MEDIA

Transnational Families and Polymedia

Mirca Madianou and Daniel Miller

ROUTLEDGE

MIGRATION AND NEW MEDIA

'An exemplary and groundbreaking study, with contributions to theory and our understanding of polymedia in everyday life, this stands out as an extraordinary read on the technology of relationships.' – Zizi Papacharissi, *University of Illinois-Chicago, USA*

'This fascinating, richly detailed book investigates the role that fluency across multiple digital platforms plays in enabling mothering and caring to be sustained at a distance. A genuine breakthrough.' – Nick Couldry, *Goldmiths, University of London, UK*

How do parents and children care for each other when they are separated because of migration? The way in which transnational families maintain long-distance relationships has been revolutionised by the emergence of new media such as email, instant messaging, social networking sites, webcam and texting. A migrant mother can now call and text her left-behind children several times a day, peruse social networking sites and leave the webcam on for 12 hours, achieving a sense of co-presence.

Drawing on a long-term ethnographic study of prolonged separation between migrant mothers and their children who remain in the Philippines, this book develops groundbreaking theory for understanding both new media and the nature of mediated relationships. It brings together the perspectives of both the mothers and children and shows how the very nature of family relationships is changing. New media, understood as an emerging environment of polymedia, have become integral to the way family relationships are enacted and experienced. The theory of polymedia extends beyond the poignant case study and is developed as a major contribution for understanding the interconnections between digital media and interpersonal relationships.

Mirca Madianou is Senior Lecturer in Media and Communication at the University of Leicester, UK. She is the author of *Mediating the Nation* and several articles on the social consequences of the media.

Daniel Miller is Professor of Material Culture at the Department of Anthropology, University College London, UK. His most recent books include *Tales from Facebook* and *Digital Anthropology* (edited with Heather Horst).

Anthropology/Media

MIGRATION AND NEW MEDIA

Transnational families and polymedia

Mirca Madianou and Daniel Miller

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2012
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada
by Routledge
711 Third Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
Madianou, Mirca.
Migration and new media : transnational families and polymedia / Mirca Madianou and Daniel Miller.
p. cm.
Includes bibliographical references and index.
1. Foreign workers, Philippine–Family relationships–Great Britain. 2. Women foreign workers–Family relationships–Great Britain. 3. Children of foreign workers–Family relationships–Philippines. 4. Communication in families–Philippines. 5. Interpersonal communication–Technological innovations–Social aspects–Philippines. 6. Communication, International–Technological innovations–Social aspects–Philippines. I. Miller, Daniel, 1954– II. Title.
HD8398.F55.M33 2011
331.40941–dc23
2011022759

ISBN: 978–0–415–67928–2 (hbk)
ISBN: 978–0–415–67929–9 (pbk)
ISBN: 978–0–203–15423–6 (ebk)

Typeset in 10/12.5pt Bembo
by Graphicraft Limited, Hong Kong

CONTENTS

1

INTRODUCTION

Within the past few years a revolution has been taking place, one with huge consequences, but so far subject to only limited systematic research. While there are many studies of globalisation and migrant transnationalism, few have addressed the consequence that probably matters most to those involved, which is the separation of families. Specifically, how do parents and children care and look after each other when they live in different countries for many years separated because of migration? Although transnational families are not new, they are becoming increasingly common. Furthermore this type of separation now often involves mothers and their children as a consequence of the feminisation of migration, partly fuelled by the insatiable demand for care and domestic workers in the developed world. The dramatic change which has revolutionised the way in which families maintain long-distance communication, is the emergence of a plethora of internet- and mobile phone-based platforms such as email, instant messaging (IM), social networking sites (SNS) and webcam via voice over internet protocol (VOIP). These new media have engendered the emergence of a new communicative environment, which we will call 'polymedia'. This book is dedicated to the understanding of this new type of 'connected transnational family' which is the result of the convergence of these two phenomena: migrant transnationalism and the explosion of communicative opportunities afforded by new media.

This book makes both a substantive and theoretical contribution to the understanding of these profound, parallel developments of family separation and transnational communication that are shaping our contemporary worlds. We believe that to understand these transformations we cannot and should not separate them as, on the one hand, a study of the media, and on the other hand, an enquiry into what it means to be a migrant, or a mother. Our understanding will be much enhanced if we study media situated in the context of what it means to be a transnational mother in this environment of polymedia. As a result, this book contains not just a theory of polymedia, but also a theory of mediation in which we consider in general terms how relationships and media are mutually shaped. We do so by drawing on a long-term ethnographic study of prolonged separation between transnational Filipino migrant mothers based in London and Cambridge and their (now adult) left-behind children in the Philippines. No other country exemplifies the phenomenon of 'distant mothering' as clearly as the Philippines with over 10 per cent of its population working overseas, the majority of whom are women with children left behind. The Philippines is also at the forefront of globalisation in terms of its appropriation of new media platforms, notably mobile phones, the consequences of which have already been documented, especially with regard to the public sphere (Castells *et al.*, 2006; Pertierra *et al.*, 2002; Rafael, 2003). More than 10 million Filipino children are officially estimated to be left-behind, most of whom see their migrant parents only once every two years. Given that such visits are even less frequent for families of undocumented migrants, it is evident that such parent–child relationships have become increasingly dependent on the available communication media. We argue that focusing on this case of prolonged separation and intense mediation helps to bring to light and crystallise aspects of both parts of this equation: a better understanding of the consequences of new media, and an insight into the very nature of parent–child relationships. Starting from this case of accentuated separation and mediation, we then move on to develop a new theory of polymedia and of mediated relationships which, we argue, can have a wider applicability. The book is equally driven by the aim to make an original contribution to the migration literature as well as to develop a theoretical understanding of digital media, distant love and the nature of mediated relationships. It also follows Stafford (2000) in arguing that understanding separation is a route towards understanding the basis of human relatedness, autonomy and dependence and thereby the very nature of relationships.

One of the book's arguments is that although information and communication technologies (ICTs) do not solve the problems of separation within families, they do contribute to the transformation of the whole experience of migration and parenting. For example, it is telling that the opportunities for cheap and instant communication feature strongly in migrant mothers' justifications regarding their decisions to migrate and to settle. However, the fact that ICTs can potentially contribute, even if indirectly, to the shaping of migration patterns is not to say that the communication is necessarily successful. In fact, we will show how the perpetual contact they engender can often increase rupture and conflict between parents and children. The only way this becomes clear is through our transnational approach to research, which involved working with both the migrant mothers and subsequently the left-behind children of these same mothers whom we interviewed back in the Philippines. In the book we both demonstrate and interpret a discrepancy between the mothers' and children's accounts. While for the mothers new communication technologies represent welcome opportunities to perform intensive mothering at a distance and to 'feel like mothers again', for their young adult children such frequent communication can be experienced as intrusive and unwanted, although this often depends on specific issues such as the age of the children at the time the mothers left and the nature of the media available to them.

In addition, this book aims to make a wider theoretical contribution by developing a theory of polymedia and a theory of mediated relationships. The theory of polymedia emerged through our need to develop a framework for understanding the rapidly developing and proliferating media environment and its appropriation by users. Although our analysis of communication technologies begins by investigating the affordances (Hutchby, 2001) and limitations of each particular medium, technology or platform, our discussion of the emergence of a new environment of proliferating communicative opportunities that is polymedia shifts the attention from the individual technical propensities of any particular medium to an acknowledgement that most people use a constellation of different media as an integrated environment in which each medium finds its niche in relation to the others. We will also argue that, as media become affordable and once media literacy[1] is established and continues to develop, the situation of polymedia amounts to a re-socialising of media itself, in which the responsibility for which medium is used is increasingly seen to depend on social and moral questions rather than technical or economic parameters.

If the term 'polymedia' recognises the importance of the human context for media use, this leads the way to our final chapter where we are able to bring this theory into alignment with the theorisation of relationships to create a theory of mediated relationships, which builds upon prior theories of mediation in media studies (Couldry, 2008; Livingstone, 2009b; Silverstone, 2005), but is here combined with debates about kinship, religion and mediation in anthropology (e.g. Eisenlohr, 2011; Engelke, 2010). In this final theoretical chapter we demonstrate that the key to understanding mediated relationships is not to envisage them as simply a case of how the media mediates relationships. Rather we start from our theory of relationships which demonstrates that all relationships are intrinsically mediated and that we can understand the impact of the media only if we first acknowledge this property of the relationship.

This book exemplifies the benefits of giving equal weight to relationships, media, ethnography and theory. But it is also sensitive to the context of its own case study, to the stories of suffering, separation, loss and also empowerment and love that make this more than just grounds for delineating such academic terrain. We have focused this volume just as much on the need to convey these stories and the background in the political economy of global labour and its impact especially on migrant women and their left-behind children.

The rest of this chapter will review the three key literatures which underpin this study, namely, global migration and transnational families; new media, consumption and transnational communication, and finally motherhood. We will end the chapter by providing an overview of the whole book.

Global migration and transnational families

Families whose members are temporally and spatially separated because of work are nothing new. Thomas and Znaniecki's classic *The Polish Peasant in Europe and*

America (1996) is a riveting account of the early twentieth-century migration to the US partly told through the letters that sustained these long-distance relationships between separated family members. The recent intensification of global migration and, crucially, the increasing feminisation of migration, have brought about a new type of transnational family where women seek employment in the global north, leaving their children behind. Transnational motherhood (Hondagneu-Sotelo and Avila, 1997), precisely because it challenges entrenched and often ideological views about the role of mothers and the value of children (see also Zelizer, 1994), has largely been seen as one of the hidden injuries of globalisation: the high social cost the developing world must pay for the increased income through remittances which keep the economies of the global south afloat.

The impact on left-behind families and the relationship people maintain to their countries of origin have been a relatively recent focus of attention, perhaps because for so long the migration literature focused on questions of assimilation and integration in the host societies (Vertovec, 2009: 13). An influential approach for understanding transnational families has been the 'care chains' approach (Hochschild, 2000; Parreñas, 2001) and the related notion of 'care drain' affecting developing countries which experience a 'care deficit' by exporting their mothers and care workers (Hochschild, 2000; Widding Isaksen *et al.*, 2008). The work of Parreñas (2001) on Philippine migration has acquired paradigmatic status in exemplifying the connections between different people across the world based on paid or unpaid relationships of care. The concept of a global care chain has particular poignancy because of the way this is refracted in the impact upon left-behind children. The paradigmatic case is where a Filipina woman from Manila spends much of her life looking after a child in London, using part of her wages to employ a Filipina from a village to look after her children in Manila. This woman in turn uses part of her urban wages to pay someone else in her village to look after her own children. These images of a global care chain are powerful representations of the larger inequalities of contemporary political economy.

There exists a corresponding debate at a popular level within the Philippines itself with regard to the impact of migration upon parent–child relationships. Critical to our fieldwork was a film called *Anak* (the word in Tagalog for 'child') which portrays the extreme example of a mother who feels she has sacrificed herself for her children by taking on domestic work in Hong Kong. But during her absence, her son drops school grades and loses his scholarship, and her daughter falls into a life of assorted vices including smoking, drinking, drugs and abusive boyfriends leading to an abortion. The film is dominated by the relationship between the mother and the sullen and resentful daughter who blames all her woes on being abandoned by her mother. This was a hugely popular film in the Philippines. It was directed by R. Quintos and starred Vilma Santos, a well-known actress and politician, now the mayor of a major town. We often started our discussion with the children by asking for their reaction to the film, which was easier to broach than immediately discussing their own childhood. So in addition to any academic debates, we also have to be aware of the way these issues are constantly appraised within the Philippines itself.

In the academic literature, gender has been understood as being key to understanding dynamics in transnational families. Parreñas (2005a) in her study of Filipino left-behind children noted that when mothers migrate they are expected to perform the caring and emotional work typically associated with their maternal role, but also to take on the traditional male breadwinning role. Globalisation and female migration have not reversed, nor even challenged traditional gender roles and hierarchies. This finding is also shared by Hondagneu-Sotelo and Avila (1997) in their study of Latina transnational mothers in California as well as Fresnoza-Flot (2009) in her research with Filipina migrants in Paris. Hondagneu-Sotelo and Avila (1997: 562) argue that female migration has not replaced caregiving with breadwinning definitions of motherhood, but rather has expanded 'definitions of motherhood to encompass caregiving from a distance and through separation'. For Pessar (1999), any advances by women's breadwinning capacity are cancelled by the fact that female migrants are overwhelmingly employed in the care and domestic sector, thus preserving patriarchal ideologies. However, McKay (2007) and Pingol (2001) observed a different gendered division of domestic work in the Philippine region of Northern Luzon.

The political economy of care and the feminist critique on which the care chains approach is based have made significant contributions to the literature on migration, with their emphasis upon the economic motivations for emigration. However, the focus of the care chains approach on structural factors does not acknowledge the empowering potential of migration for women and does not grant much agency to migrants themselves in determining their own trajectory (McKay, 2007; Silvey, 2006; Yeates, 2004). The care chains approach also assumes a normative and universal perspective of biological motherhood which should be performed in a situation of co-presence (actually living together in the same household). What the more ethnographically based studies such as McKay (2007) and Aguilar *et al.* (2009) demonstrate is that both the global feminist discourse employed by Parreñas (2001), and also globalised ideas about women's responsibilities (which are found in the Hollywood-style melodrama that clearly influenced the film *Anak*) have to be complemented by grounded study within the Philippines, which may reveal very different and more nuanced expectations about mother–child relationships.

Mothers themselves are subject to competing discourses about the moralities of their own actions. In such circumstances it seemed vital to recognise the migrant women's own perspective, particularly when the research agenda concerns sensitive and emotive issues such as family separation. In our research we have adopted an ethnographic approach which recognises migrants as reflexive subjects, albeit ones positioned in structures of power. For example, crucial for understanding the relationships and communication between mothers and left-behind children is the analysis of the context of migration, including the reasons why women migrate in the first place. The bottom-up ethnographic perspective followed here can uncover the contradictory and perhaps less socially acceptable motivations for migration and cast light on the processes through which women negotiate their various roles, identities and relationships. This is an approach followed by

Constable (1999) in her work with Filipina migrant workers in Hong Kong, where she focused on the ambivalent narratives of return amongst her participants. Such accounts of the motivations for migration and settlement often highlight personal reasons which are not captured by more top-down perspectives such as that of the care chains with its emphasis on the role of the state and the political economy of care.

In Chapter 3 we build upon Parreñas (2001: 27) who developed an intermediate level analysis combining a bottom-up perspective with the macrostructural approach of political economy of labour migration (Sassen, 1988). This allowed Parreñas to identify a range of 'hidden motivations for migration' which extend beyond the well-rehearsed and socially accepted reasons, which are usually economic. For example, Parreñas observed that personal reasons including the breakdown of a relationship, domestic abuse and extramarital affairs, constitute a significant motivation for women's migration (2001: 62–69), often in conjunction with other well-documented economic and political reasons. However, our work suggests that migrants do not always articulate the contradictions (what Parreñas calls the 'dislocations of migration' [2001: 23]) in their narratives. Rather, often the discrepancy between their own accounts (which often draw on well-rehearsed public discourses about what constitutes good mothering and a good reason to go abroad) and their actual practices, points to the contradictions and ambivalence that is part of the project of migration. To unearth such discrepancies one needs the long-term and in-depth involvement of ethnography. Migrant women occupy simultaneously different and often contradictory subject positions: breadwinners and caregivers; devoted mothers and national heroines; global consumers and exploited workers. Our ethnographic perspective documents how they negotiate these conflicting identities both discursively and through practices.

Although, as we noted earlier in this section, research on transnational families is part of the transnational turn within migration studies, it is perhaps ironic that one still encounters a degree of 'methodological nationalism' (Wimmer and Glick-Schiller, 2002) within such scholarship. It is as if researchers cannot escape the 'assumption that the nation/state/society is the natural social and political form of the modern world' (Wimmer and Glick-Schiller, 2002: 301). Although it would be foolish to entirely repudiate the relevance of the nation-state in the analysis of migrant transnationalism, it seems that one way of overcoming the straightjacket of methodological nationalism is to actually conduct research transnationally. Our research has benefited from this comparative, multi-sited perspective. By focusing on the relationships between migrant mothers based in the UK *and* their left-behind children in the Philippines we have 'followed the thing' through a multi-sited ethnography (Marcus, 1995). The comparisons – and contradictions – between the mothers and the children's perspectives lie at the heart of this book. We came to recognise that we would have written an entirely different book if we had concentrated on migrant mothers only, or on their children. Transnationalism is all about relationships, and following them (rather than assuming them) is one way of dealing with the perils of methodological nationalism.

Transnational communication and new media

For transnational families who are reunited on average every two years,[2] new media are essential for keeping in touch. Dependence on new media is exacerbated in the case of irregular migrants who often do not see their families for longer periods (in our sample the longest period without a visit was 13 years; for similar observations see also Fresnoza-Flot, 2009). In such cases new communication technologies become the only means through which migrant mothers can maintain a relationship with their children. Given this almost extreme dependency, it is perhaps surprising that new media have not received much attention in the literature of migrant transnationalism, although studies have highlighted the more general importance of the mobile phone as a social resource in the lives of migrants (see Thompson, 2009). Most academic writing on new media and migration has looked at the important questions of identity and integration (Gillespie *et al.*, 2010) and the political implications for diasporic and national populations (Brinkerhoff, 2009; for a review see Siapera, 2010). Although this literature has been very useful and influential, it does not address the urgent question of sociality and intimacy in a transnational context (although see Horst, 2006; Miller and Slater, 2000; and Wilding, 2006), while the focus on the rather bounded concept of identity does not always capture the dynamic nature of transnational processes (Glick-Schiller *et al.*, 1992; Madianou, 2011).

In the context of Philippine migration, Parreñas observed that among separated Filipino families mobile phones actually tie migrant women to their traditional gender roles (Parreñas, 2005b), echoing North American studies about mobile phone use and the spillover of the domestic into the professional sphere (Chesley, 2005; Rakow and Navarro, 1993). Apart from gender inequalities, Parreñas also argues that the political economic conditions of communication determine the quality of transnational intimacy and family life (Parreñas, 2005b), as families without access to the internet or even a landline are deprived of care and emotional support. In the next chapter we shall acknowledge these stark asymmetries both between the communications infrastructure of the Philippines and the UK and within the Philippines. But although a political economic analysis has to inform our understanding of transnational family communication, it cannot fully account for the dependency of such families on digital media and the mutual shaping of technologies and relationships.

The greatest challenge of studying new media in the context of transnational family relationships is that the technologies themselves are constantly changing and research often seems to be chasing a moving target of technological developments and innovative appropriations on the part of the users. Each new mode of communication seems to become what Vertovec (2004) described for cheap international cards, that is 'the social glue of transnationalism', with examples provided by Wilding (2006) on email, or Uy-Tioco on texting (2007). This is reflected in our own studies. When we began our fieldwork three years ago, transnational family communication was often mainly centred upon one medium such as telephone calls or email, each with its own affordances and limitations. It was often possible to see the consequences of that particular type of communication on the relationships in question. However, gradually, and certainly over the past couple of years, we noticed a shift towards a situation of multiple media. Relationships, increasingly, do not depend on one particular technology, but on a plurality of media which supplement each other and can help overcome the shortcomings of a particular medium. People can also take advantage of these different communicative opportunities in order to control the relationship. So, for example, if they want to avoid confrontation they do not call but send an email. This is what led us to consider polymedia with a focus on the social and emotional consequences of choosing between a plurality of media rather than simply examining the particular features and affordances of each particular medium (see also Baym, 2010; Gershon, 2010).

Although we recognise that this new environment of communicative opportunities is not yet a reality for everyone in the Philippines or even in the UK, it already represents a qualitative shift in the way technologies mediate relationships. This is why we felt the need for this new term to allow us to describe the situation. Although the term 'media ecology' could be an alternative, it is concerned with the wider systems of communication such as transport, or issues of usage such as politics and health (Slater and Tacchi, 2004), while we wanted a term that will highlight the unprecedented plurality and proliferation of media. 'Multimedia', on the other hand, is now an established term with a very different meaning (a situation where several different forms of media are being used simultaneously and in direct relationship to each other, for instance using instant messaging on social networking sites) and it would therefore be confusing to use that term. 'Multichannel', or 'multi-platform' might be closer to what we wish to describe, although choosing either term would force us to prioritise either the terms 'platform' or 'channel' when in fact our findings suggest that such technological hierarchies are not particularly meaningful to users. This is why we chose 'polymedia' as a new term to describe the new emerging environment of proliferating communicative opportunities.

It may seem that the term 'polymedia' merely acknowledges the plethora of different media that are now available, but the point we wish to make is both more profound and closer to the heart of social science. Our argument will be that this growth of diverse media is crucially linked to changes in their pricing structure as well as in users' media literacy (Livingstone, 2004), and it is the combination of these factors that transforms the relationship between people and media. Previously people would assume that the choice of media was dictated by issues of availability of technology or price, and were constrained from extending their inferences from the choices other people made. But the word 'polymedia' will be used to consider much more generally how media are socialised, which is why it then leads on to a subsequent theory of mediation. Both draw upon a number of theoretical developments such as the rich tradition of consumption and domestication of ICTs (Berker *et al.*, 2006; Miller and Slater, 2000; Silverstone and Hirsch, 1992) and the theory of mediation (Chouliaraki, 2006; Couldry, 2008 and 2012; Eisenlohr, 2011; Livingstone, 2009b; Madianou, 2005 and 2012b; Miller, in press; Silverstone, 2005).

Historically, mediated interaction was understood as being inferior compared to the golden standard of face-to-face. This was mainly due to the reduced amount of symbolic cues (for example, lack of visual cues in a letter, or telephone communication) which gave rise to ambiguities and potential misunderstandings (Baym, 2010: 51–54; Thompson, 1995: 84). Also problematic was the perceived lack of norms to regulate mediated interaction which also had the potential to amplify conflict (through the case of 'flaming') (Baym, 2010: 55). Recent studies on the social shaping of technologies (MacKenzie and Wajcman, 1999; Wajcman *et al.*, 2008), domestication (Berker *et al.*, 2006; Miller and Slater, 2000) and mediation (Couldry, 2008; Livingstone, 2009b; Madianou, 2005; Silverstone, 2005) have shown that mediated interactions are more complex than that and that society and relationships are mutually constitutive. Similarly, polymedia aims to contribute to this academic discussion by showing how users can overcome the limitations of any particular medium by choosing an alternative in order achieve their communicative intents and to assume control over their relationships. We should stress, however, that we are not implying that media power is becoming redundant in a situation of polymedia. On the contrary, power is a recurrent theme in this volume and will be analysed as being present in both the social and family contexts (family relationships are asymmetrical) and the political and economic contexts of migration and telecommunications.

All of these are brought together in our final chapter, which culminates in a theory of mediation which can be traced back to the early days of media and communications research when Lazarsfeld and Merton wrote in 1948 that research ought to try to understand the effects of the sheer presence of media institutions on society. Silverstone developed this notion of mediation as:

> a fundamentally dialectical notion which requires us to understand how processes of communication change the social and cultural environments that support them as well as the relationships that participants, both individual and institutional, have to that environment and to each other. At the same time it requires a consideration of the social as in turn a mediator: institutions and technologies as well as the meanings that are delivered by them are mediated in the social processes of reception and consumption (Silverstone, 2005: 3).

In this volume the term 'mediation' applies just as much to the question of what is a social relationship as to the question of what is a medium. The situation of transnational mothering raises huge issues of what the very terms 'mother' and 'child' mean. When we come to the analysis of our research material it will become clear that this extreme case actually throws light on the question of what is a mother as it applies to any situation including that of co-presence because a social relationship is already a form of mediation. A mother is both a normative concept – the ideal as to what a mother should be – and the experience of actually being, or having, a mother. As this book unfolds it will show why any further development of a theory of mediation as applied to media is best achieved through equal attention to the theory of mediation as to the relationship. This point leads directly to our third discussion of the literature, which starts to open up this question of what we mean when we use the term 'mother'.

Transnational motherhood: normativity and ambivalence

The last section suggests that in order to assess the impact of ICTs on the relationships between transnational mothers and their left-behind children, we need to pay just as much attention to the issue of motherhood as to an appreciation of the media. In the next chapter we will provide a discussion of the Filipino idioms of family, motherhood and childhood. The present discussion is a more general reflection on motherhood which we argue is indispensable to the understanding of mother–child relationships and their mediation. One of the reasons such a clarification of terms and theoretical baggage is crucial is because motherhood is a constant trope in ideological debate. Moral panics regularly erupt about what constitutes good, or 'good-enough' mothering (Winnicott, 1971), feeding into questions and often translated into policy regarding mothers' employment and identities (Riley, 1983; Rutter, 1981; Smart and Neale, 1999). Even though there is an increasing recognition of the changing nature of family and the plurality of parenting arrangements (Golombok, 2000) – from single mothers and working mothers to stay-at-home mothers and from heterosexual mothers to lesbian mothers and so on – there is widespread assertion, even among feminists, that parenting needs to take place in a situation of co-presence. This is nothwithstanding anthropological accounts of many regions, such as the Caribbean, where the nuclear co-present family has never been the norm (for a classic account see Clarke, 1957). In this climate mothers leaving their children to pursue their own ambitions are quickly branded 'bad mothers' (Jackson, 1994; for a discussion of the Philippine case see Parreñas, 2008: 22–39). Transnational mothering disrupts this normative notion of co-present parenting. This is the dominant discourse that we now see reflected in the film *Anak* discussed earlier. Our purpose in the next few paragraphs is to clarify our analytical tools that will help us understand the changing nature of mother–child relationships in the context of separation and mediation.

Being a mother is defined by being in a particular relationship. As Miller (1997) has shown in his paper 'How infants grow mothers in North London', the development of the child as a new being is equally reflected in the process, much more commonly taken for granted, by which a female becomes a mother. This disrupts the dominant literature which is driven by a concern to examine the impact of maternal behaviour on children's development. Feminist critique has for a long time identified such one-sided emphasis in psychoanalysis and developmental psychology (Hollway and Featherstone, 1997; Parker, 2005: 15–18). Even

Winnicott's (1975) essay 'Hate in the countertransference', which is now considered a classic text on maternal ambivalence, is concerned with its impact on the baby's development. To simply see the mother from the point of view of her child's needs would be tantamount to her infant's own narcissism: seeing the mother as merely an extension of the infant's needs. Parker (2005: 18) refers to 'maternal development' to contest the one-sided emphasis on child development, thus acknowledging that mothers, just like their children, are changing also as part of the challenges of the experience of motherhood. This is part of Parker's efforts to theorise mothers, rather than treat them as empty vessels to be filled with their children's needs and desires. Our research contextualises motherhood in the wider lives of these Filipina women, subjects with multiple identities and needs, that can become manifest as ambivalence. Although most writing on ambivalence is located within psychoanalysis (Hollway and Featherstone, 1997; Parker, 2005), we feel it is essential to recognise these same issues within an ethnographic encounter that can equally expose what Hays (1997) terms the 'cultural contradictions of motherhood'.

Our evidence will show that ambivalence is particularly relevant to the experience of migration. For mothers with left-behind children, migration as deterritorialisation can exacerbate such maternal ambivalence. We regard ambivalence as a normal state for many mothers (Hays, 1997), who must negotiate contradictory roles as workers and mothers, but equally the ideal freedoms posed by feminism contradicted by the constraints and re-gendering created by motherhood (Miller, 1997). For migrant mothers such negotiation is more challenging because work (in the UK) and mothering (in a transnational space) are spread across different countries and continents, leading to a situation of 'accentuated ambivalence' (Madianou, 2012a). In Chapters 3 and 5, which aim to illustrate the contours of transnational mothering from the bottom up, we will explore the ways in which mothers negotiate this ambivalence and the role that ICTs play in this process.

This insistence upon acknowledging the perspective of mothers need not be opposed to the perspective of their children, which for us would be an abnegation of our understanding of the constitution of both mother and child as a relationship. According to Hollway (2001) the literature on motherhood and child development seems to have been marked by a certain dualism between those perspectives which see mothers as objects of their babies' and their families' needs (rather than people in their own right) and the feminist critique which sees women as subjects and active agents. Hollway (2001) proposes an alternative perspective of intersubjectivity as part of the attempt to examine mothers' and children's needs *in tandem*. Accordingly, here we directly juxtapose the points of view of both mothers and children. We were able to accomplish this because methodologically we sought to interview not only the mothers in the UK, but then with their permission their left-behind children – now young adults – in the Philippines. In total we were able to pair 20 mothers and children, but our wider sample contains many more mothers and children who were not 'paired up' (see Appendix for a detailed discussion of our sample and overall method).

Studying Facebook via Data Extraction: The Netvizz Application

Bernhard Rieder
University of Amsterdam
Turfdraagsterpad 9
1012TX Amsterdam
rieder@uva.nl

ABSTRACT

This paper describes Netvizz, a data collection and extraction application that allows researchers to export data in standard file formats from different sections of the Facebook social networking service. Friendship networks, groups, and pages can thus be analyzed quantitatively and qualitatively with regards to demographical, post-demographical, and relational characteristics. The paper provides an overview over analytical directions opened up by the data made available, discusses platform specific aspects of data extraction via the official Application Programming Interface, and briefly engages the difficult ethical considerations attached to this type of research.

Author Keywords
research tool, social networking services, Facebook, data extraction, social network analysis, media studies

ACM Classification Keywords
J.4 Social and Behavioral Sciences

INTRODUCTION

In October 2012, Facebook announced that it had reached the symbolic number of one billion monthly active users. [4] This arguably makes it one of the biggest media organizations in the history of humankind, contested only by Google's collection of services in terms of daily worldwide audience size and engagement. Traditional corporations dwarf these massive Internet companies when it comes to the size of their workforce – Facebook employed a mere 4500 people at the end of 2012 – but the sheer number of "[p]eople [who] use Facebook to stay connected with friends and family, to discover what's going on in the world, and to share and express what matters to them" [4] is simply gigantic. It is no wonder, then, that researchers from many areas of the human and social sciences have moved quickly to study the platform: a recent review article [19] identified 412 peer-reviewed research papers that follow empirical approaches, not counting the numerous publications employing conceptual and/or critical approaches. While traditional empirical methods such as interviews, experiments, and observations are widely used, a growing number of studies rely on what the authors call "data crawling", i.e. "gleaning information about users from their profiles without their active participation" [19]. This paper presents a software tool, Netvizz, designed to facilitate this latter approach.

Research methods using software to capture, produce, or repurpose digital data in order to investigate different aspects of the Internet have been used for well over a decade. Datasets can be exploited to analyze complex social and cultural phenomena and *digital methods* [12] have a number of advantages compared to traditional ones: advantages concerning cost, speed, exhaustiveness, detail, and so forth, but also related to the rich contextualization afforded by the close association between data and the properties of the *media* (technologies, platforms, tools, websites, etc.) they are connected with; data crawling necessarily engages these media through the specifics of their technical and functional structure and therefore produces data that can provide detailed views of the systems and the use practices they host. The study of social networking services (SNS) like Facebook, however, introduces a number of challenges and considerations that makes the scholarly investigation of these services, their users, and the various forms of content they hold significantly different from the study of the open Web. This paper discusses some of the possibilities and difficulties with the data crawling approach applied to Facebook and introduces a tool that allows researchers to generate data files in standard formats for different sections of the Facebook social networking service without having to resort to manual collecting or custom programming. I will first introduce some of the approaches to data extraction on SNS, in order to situate the proposed tool. I will then introduce the Netvizz application and provide a number of short examples for the type of analysis it makes possible. Before concluding, I will discuss two further aspects that are particularly relevant to the matter at hand: research via Application Programming Interfaces (API) and the question of privacy and research ethics. While this paper contains technical descriptions, it is written from a media studies perspective and therefore focuses on aspects most relevant to media scholars.

WebSci'13, May 2–4, 2013, Paris, France.
Copyright 2013 ACM 978-1-4503-1889-1....$10.00.

STUDYING FACEBOOK THROUGH DATA EXTRACTION

The study of Internet platforms via data extraction has seen fast growth over the last two decades and the recent excitement around the concept of *big data* seems to have added additional momentum to efforts going into this direction. [9] For researchers from the humanities and social sciences, the possibility to analyze the expressions and behavioral traces from sometimes very large numbers of individuals or groups using these platforms can provide valuable insights into the arrays of meaning and practice that emerge and manifest themselves online. Besides merely shedding light on a "virtual" space, supposedly separate from "real life", the Internet can be considered as "a source of data about society and culture" [12] at large. The promise of producing *observational* data, i.e. data that documents what people do rather than what they say they do, without having to manually protocol behavior, expressions, and interactions is particularly enticing to researchers. SNS in general, and the gigantic Facebook platform in particular, can be likened, on a certain level, to observational devices or even to experimental designs: the "captured" data are closely related to meticulously constructed technical and visual forms – functionalities, interfaces, data structures, and so forth – that function as "grammars of action" [1], enabling and directing activities in distinct ways by providing and circumscribing possibilities for action and expression. Even if the design of this large-scale social experiment is specified neither by nor for social scientists and humanists, the delineated and parametered spaces provided by SNS confer a controlled frame of reference to gathered data. No wonder that Cameron Marlow, one of the research scientists working at Facebook considers the service to be "the world's most powerful instrument for studying human society" [16]. In order to better understand how such data can be gathered, a short overview of existing approaches is indispensable.

Existing Approaches

The already mentioned review paper [19] distinguishes five categories of empirical Facebook research: descriptive analysis of users, motivations for using Facebook, identity presentation, the role of Facebook in social interactions, and privacy and information disclosure. It is not difficult to see how approaches gathering data from or through the platform can be useful for each of these areas of investigation. The question, then, is what data can actually be accessed and how this is to be done, considering that the particular technique chosen has important repercussions for the scope of what can be realistically acquired.

One can largely distinguish two general orientations when it comes to collecting digital data from SNS through software-based tools: first, researchers can recruit participants, through Facebook itself or from the outside, and gather data by asking them to fill out questionnaires, often via so called Facebook applications[1]. [11] While this method certainly differs from traditional ways of recruiting participants in terms of logistics and sampling procedures, it is not fundamentally different from online surveying in general.[2] Second, data can be retrieved in various ways from the pools of information that the Facebook platform *already* collects as part of its general operation. This latter approach, which is the focus of this paper, is fueled by data derived from both sides of the distinction Schäfer makes between "implicit and explicit participation" [14], referring to the difference between information and content deliberately provided by users, e.g. by filling out their profiles, and the data collected and produced by logging users' actions in sometimes minute detail. While Facebook members share content, write messages, and curate their profiles, they also click, watch, read, navigate, and so forth, thereby providing additional data points that are stored and analyzed. Because these activities revolve around elements that have cultural significance – liking a page of a political party is more than "clicking" – these data are not simply behavioral, but allow for deeper probing into *culture*. For research scholars, there are three ways by which to gain access to these data, with significant differences between approaches in terms of technical requirements and institutional positioning:

Direct database access to the company's servers is reserved to in-house researchers or cooperation between a SNS and a research institution. [17] Certain companies also make data "donations", for example Twitter deciding to transfer its complete archive to the Library of Congress, albeit with a significant delay. The data made accessible in these ways are generally very large and well structured, but often anonymized or aggregated. Partnering with a platform owner is certainly the only (legal) way to gain access to *all* collected data, at least in theory.

Access through sanctioned APIs makes use of the machine interfaces provided by many Web 2.0 services to third-party developers with the objective of stimulating application development and integration with other services in order to provide additional functionality and utility to users. These interfaces also provide well-structured data, but are generally limited in terms of which data, how much data, and how often data can be retrieved. Conditions can vary significantly between services: in contrast to Twitter, for example, Facebook is quite restrictive in terms of what data can be accessed, but imposes few limits on request frequency. Companies also retain the right to modify or close their data interfaces, which can lead to substantial problems for researchers.

[1] A Facebook application is a program that is provided by a third-party but integrates directly into the platform.

[2] One should note that studies using questionnaires on Facebook often access profile data as well.

User interface crawling can be done manually, but usually employs so-called *bots* or *spiders* that read the HTML documents used to provide graphical interfaces to users, either directly at the HTTP protocol level or via browser automation from the rendered DOM.[3] [8] These techniques can circumvent the limitations of APIs, but often at the price of technical and legal uncertainties if a platform provider's permission is not explicitly granted. In the case of Facebook, bot detection mechanisms are in place and suspicious activity can quickly lead to account suspension.

If performed on a large scale, all of these approaches require either custom programming or considerable amounts of manual work. The focus points and requirements for research and teaching do, however, bear marks of resemblance and Facebook itself is designed around a limited number of functionalities or "spaces". One can therefore argue that general-purpose tools may be envisioned that provide utility to a variety of research projects and interests. Several such *data extractors* targeting Facebook have been developed over the last years, invariably using sanctioned APIs for data gathering. These tools generally export data in common formats and they focus on specific sections of the platform – partly by choice, partly due to limitations imposed by the platform itself. Their goals are also similar: to lower the technical and logistical requirements for empirical research via data analysis in order to further the ability of researchers to study a medium that unites over a billion users in a system that is essentially conceived as a walled garden. In what follows, I describe the Netvizz application[4], a tool designed to help research scholars in extracting data from Facebook.

Similar Work

The enormous success of Facebook has prompted the emergence of a large number of analytics tools for marketing purposes, which often focus on *pages*, the section of Facebook that brand communication and consumer relations rely on, due to their public showcase character. Because these tools are generally built for monitoring marketing campaigns, they target page *owners* rather than researchers interested in studying a page. For this reasons – and the sheer number of tools available – I will leave these applications to the side.

There are, however, two tools that function as general-purpose data extractors for researchers studying Facebook. *NameGenWeb*[5] originated at the Oxford Internet Institute and provides the possibility of exporting a user's friendship network, i.e. all of the user's friends, the friendship connections between them, and a wide array of variables for each user account extracted. Another application, the *Social Network Importe*r[6], a plug-in for the *NodeXL* network analysis and visualization toolkit developed by an international group of scholars, provides similar functionality for downloading personal networks, but also a means to extract extensive data from Facebook pages, including monopartite[7] networks for users and posts, based on co-like or co-comment activities, and bipartite networks combining the two in a single graph. One should also mention Wolfram Alpha's "Facebook report"[8] in this context: while it does not make raw data available, and therefore limits in-depth analytics using statistical or graph theoretical approaches, the tool provides a large number of analytical views on personal networks.

The Netvizz application provides "raw" data for both personal networks and pages, but provides data *perspectives* not available in other tools, e.g. comment text extraction; it also provides data for groups, a third functional space on Facebook. Running as a Web application, Netvizz does not require the use of Microsoft Excel on Windows like *NodeXL* and thereby further lowers the threshold to engagement with Facebook's rich data pools. The next section will introduce the application and its different data outputs in more detail.

THE NETVIZZ APPLICATION

The Netvizz application was initially developed by the author in 2009 as a practical attempt to study Facebook's API as a new media object in its own right[9] and to gauge the potential of using natively digital methods [12] to study SNS. Because of the positive reactions and high uptake, the application was developed into a veritable data extractor that provides outputs for different sections of Facebook in standard formats.[10] Before introducing the different

[3] The latter approach has become more common due to the fact that sites are increasingly using programming languages (mostly JavaScript) to assemble pages client-side rather than sending finished documents described in a markup language (mostly HTML).

[4] https://apps.facebook.com/netvizz/

[5] https://apps.facebook.com/namegenweb/

[6] http://socialnetimporter.codeplex.com/

[7] Monopartite graphs contain nodes that are all of the same kind (e.g. users). Bipartite graphs include two types of nodes (e.g. users and posts), and so forth.

[8] http://www.wolframalpha.com/facebook/

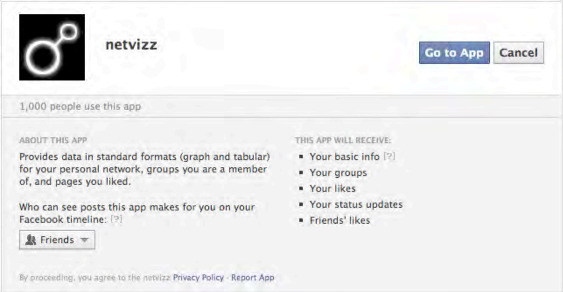
[9] APIs as *objects* of research for new media scholars are only slowly coming into view, despite their importance for the Web as data ecosystem. A separate publication will detail empirical approaches to studying APIs from a critical media studies perspective.

[10] Data formats were chosen for their generality and simplicity. Network outputs use the GDF format introduced with the GUESS graph analysis toolkit. Tabular outputs use a simple tab separated format that can be opened in virtually all spreadsheet applications and statistical packages.

features, it is necessary to briefly discuss the Facebook API and those characteristics that are relevant to research procedures and data quality.

Data Access via the Facebook API

As indicated above, Netvizz is a simple Facebook application written in PHP that runs on a server provided by the Digital Methods Initiative[11]. It is part of Facebook's app directory and can be found by typing the name into the platform's main search box. Like any other Facebook application, it requires users to log in with an existing Facebook account to be able to access any data at all.



1,000 people use this app

ABOUT THIS APP

Provides data in standard formats (graph and tabular) for your personal network, groups you are a member of, and pages you liked.

Who can see posts this app makes for you on your Facebook timeline: [?]

THIS APP WILL RECEIVE:

- Your basic info [?]
- Your groups
- Your likes
- Your status updates
- Friends' likes

By proceeding, you agree to the netvizz Privacy Policy · Report App

Figure 1. The Netvizz app permission request page.

A vast SNS that deals with intimate and potentially sensitive matters is likely to implement rather strict privacy policies and this is – to a certain extent – also the case with Facebook. The construction of the Facebook API reflects these concerns in at last four ways that are significant here:

First, every probe into the data pool is "signed" with the credentials of a Facebook user whose actual status on the platform defines the scope of which data can be accessed. For example, detailed user data can generally only be extracted from accounts a user is friends with and one has to be a member of a group to extract any data from it.

Second, users' privacy settings play a role in what data can be exported. If one user excludes another from seeing certain elements on his or her profile, an application operating with the latter's credentials will also be blocked from accessing those elements.

Third, every application is required to explicitly ask for permission to access different data elements.[12] These requests are displayed to the user when she first uses the application. Figure 1 shows the permission dialogue for the Netvizz application. While these permissions have to be given for the application to work, users can limit the data made available to applications used by their friends in their preferences.

Fourth, certain elements that are visible on the level of the user interface are not available through the API. The user view count displayed on each post in a group, for example, is (currently) not retrievable and certain data elements, such as friends' email addresses, are equally off limits by design.

While we can expect scholars using the Netvizz application to grant all the permissions[13] it asks for – it will simply not work otherwise – users' privacy settings are indeed relevant when it comes to interpreting the retrieved data: from a technical perspective, it is not possible to know whether an empty field is empty because the user has not filled in the specific data or because the privacy settings prohibit access. This must be taken into account when making assumptions on the basis of missing data. User profile data, in particular, should be handled with prudence. Other data, such as page engagement and friendship connections in personal networks and groups, can be considered robust, however.

Overview

The Netvizz application currently extracts data from three different sections of the Facebook platform:

Personal networks are considered in two different ways. First, the friendship network feature provides a simple undirected graph file where the friends of the logged user are nodes and friendship connections edges. Sex, interface language, and a ranking based on the account creation date[14] are provided for each user and counts for posts and likes can be requested as an option. Friendship networks often cluster around significant places in a user's life, e.g. geographies or institutions such as high school, university, workplaces, clubs, and so forth. Second, a bipartite "like network" can be generated that formalizes both users and liked entities (all elements already represented in Facebook's Open Graph[15] are extracted) as nodes, a user liking a page generating an edge. This network, examined via a graph analysis toolkit, will arrange both users and liked objects around cultural affinity patterns, foregrounding *post-demographic* [13] variables.

Groups can be explored in a similar fashion as friendship networks, although the API currently limits the number of users one can retrieve from a group to 5000. For larger groups, a random subset of users is provided. A second

[11] https://www.digitalmethods.net

[12] For details concerning the permission structure refer to: http://developers.facebook.com/docs/reference/login/

[13] The Netvizz application does not store or aggregate any of the extracted data in a database and the generated files are deleted in regular intervals.

[14] The unique identifiers for accounts on Facebook are numbered consecutively, which means that the lower the number, the older the account. Netvizz simply adds a ranking to the output that orders accounts by their age.

[15] For more information on how Facebook represents entities in the *Open Graph* concept modeling system, see: https://developers.facebook.com/docs/concepts/opengraph/.

feature also provides a *social* graph, but one that is based on interactions between group members through the posts sent to a group. If one user likes or comments on another user's post, a directed edge between the two users is created, each interaction adding weight to the edge.

Pages are represented as a bipartite network, with both posts (up to 999 latest posts) and users as nodes. If a user comments on or likes a post, a directed edge between user and post is created. This way, one can not only detect the most active users, but also identify the posts that produced the highest amount of engagement. The latter data are also provided in a tabular data file, ready for statistical analysis. To make content analysis easier, a third file containing user comments, grouped per post, is generated. The application allows selecting whether posts made by users should be included, in addition to posts made by the page owner.

ANALYTICAL DIRECTIONS

The two types of data files provided by Netvizz – network files and tabular files – already indicate basic directions for analytical approaches, the former allowing for the application of concepts and methods from Social Network Analysis [15] and Network Science [18], while the latter points towards more traditional statistical techniques. Before describing analytical approaches in more detail, a short comment on modes of analysis – and in particular visualization – is in order.

Analysis and Visualization

One of the reasons for choosing simple and common file formats for outputs in Netvizz was the need to compensate for the lack of an actual visual and analytical interface in the application itself. There are, indeed, a number of Facebook applications available that produce direct visual representations, generally of personal networks, which greatly facilitates the initial encounter with the data in question for researchers with little or no training in quantitative research. Because these tools are mostly visualization widgets that do not target researchers and offer little to no analytical methodology beyond the visual display itself, one of the initial intentions was to design Netvizz as a bridge between Facebook data and the various network analysis toolkits available today, such as GUESS[16], Pajek[17] or the very easy to use Gephi[18]. The last program, in particular, must be credited with significant lowering the threshold to working with network analysis and visualization. Netvizz voluntarily inscribes itself in a movement, epitomized by tools such as gephi and the work of the Amsterdam-based Digital Methods Initiative[19] and other groups, that aims at bringing data-driven analysis to a wider audience and, specifically, to an audience that includes those regions of the social sciences and humanities that have been shunning quantitative and computational methods because of the epistemological and methodological commitments often associated with quantification and formalization. Lowering the threshold to using computer-based analytical methods is therefore not simply a service to long-time practitioners, but an attempt to see in what way and how far these methods can be useful in contexts where the dominant "styles of reasoning" [7] are based on interpretation, argumentation, and speculation, and build on conceptualizations of human beings and their practices that simply cannot be formalized as easily as theoretical frameworks like behaviorism or social exchange theory.

In this context, visualization has been presented as a means to profit from the analytical capacities afforded by software without having to invest years into the acquisition of skills in statistics or graph theory. While the data provided by Netvizz can certainly be used to calculate correlation coefficients as well as network metrics, focus was put on facilitating analysis through visualization. There is, however, no need to juxtaposition mathematical and visual forms of analysis; as Figure 2 demonstrates, the latter can not only help in communicating the results provided by the former, but adds a way of relating to the data that can provide a significant epistemic surplus.

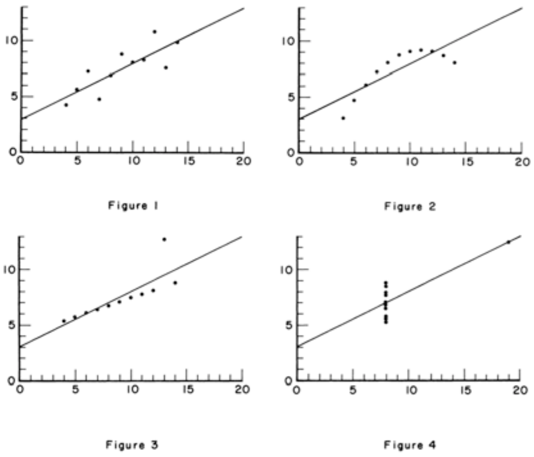


Figure 2. Four scatter-plots from [2]. They have identical values for number of observations, mean of the x's, mean of the y's, regression coefficient of y on x, equation of regression line, sum of squares of x, regression sum of squares, residual sum of squares of y, estimated standard error of bi, and multiple r2. Yet, the differences between the dataset are strikingly obvious to our eyes. Anscombe uses this example to make an argument for the usefulness of visualization in statistics beyond the communication to a larger audience.

Independently of its application to actual empirical analysis of Facebook data, Netvizz should thus be considered a pedagogical tool that can help in getting started with quantitative methodology, network analysis, and the

[16] http://graphexploration.cond.org

[17] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[18] https://gephi.org

[19] https://digitalmethods.net

required software. While one could argue that network visualizations are images and therefore intuitively accessible and "readable", there are also arguments that point into the opposite direction. It is easy to show how different graph layout algorithms highlight particular properties of a network and familiarity with a dataset can go far in helping novice users understand what is actually happening when they use software to work with graph data. Because many people are intimately familiar with their Facebook networks, they can more easily see what the software does, and what kind of epistemic surplus one can potentially derive from network analysis.

Analytical Perspectives

In actual research settings, Netvizz can provide data relevant to many different approaches and research questions. One can also consider different embeddings in the logistics of research projects: it is imaginable that a study recruits users to investigate patterns in social relations, but instead of asking them for access to their accounts, they encourage them to run the Netvizz application from their profile and share the data with the researchers. Descriptive approaches to user profiling could thus complement traditional socio-economic descriptors with *post-demographic* properties [13] in the form of like data and the *relational* data represented by friendship networks. It is worth mentioning that Netvizz uses the unique Facebook account identifiers as "keys" for nodes in the GDF format; this means that all network files can be combined to form larger networks because the same user appearing in two different files will be a single node if the networks are combined, e.g. in gephi.

The group and page features also enable or facilitate data-driven approaches to studying Facebook users and uses without requiring access to individual accounts. In the case of groups, one needs to be a member to access its data; in the case of pages, liking it is enough to make it show up in the Netvizz interface. The analytical possibilities afforded by the second perspective are explored in more detail via two short case studies in the following section, but one could classify analytical dimensions along a series of very basic questions:

Who? This concerns studies of users (profile data), their relations (friendship patterns and interactions), and the larger social spaces emerging through groups and pages.

What? For personal networks, this relates mainly to *likes*, while pages allow for an investigation into *posts*, in particular concerning media types and audience engagement.

Where? For all outputs containing information about users, interface language is provided in a comprehensive way, because users do not have the possibility to prevent applications from receiving this information. While interface language is certainly not a perfect stand-in for locality, it allows engaging the question of geography in interesting ways.

When? Temporal data is limited to pages, but here, a timestamp for each post and comment is provided, allowing for investigating page and user activity over time.

EXAMPLES

To make the provided directions for analysis more tangible, this section briefly outlines two case studies investigating the use of Facebook in political activism online, more precisely its use by the anti-Islam movements that have grown at a rapid pace, in particular since the 9/11 attacks. The first example focuses on a group and the second on a page. Both examples mobilize concepts and techniques from Social Network Analysis (SNA), which developed out of the work of social psychologists Jacob Moreno and Kurt Lewin in the 1930s and 1940s. Although its tight relationship with social exchange theory [3] has granted a certain amount of visibility to SNA, it is only the wide availability of relational data and the software tools to analyze these data that the approach has gained the popularity it enjoys today. The main tenant of SNA is to envision groups and other social units as *networks*, that is, as connected ensembles that emerge from tangible and direct connections (friendships, work relationships, joint leisure, direct interactions, etc.) rather than as social *categories* that are constructed on the bases of shared (socio-economic) properties instead of actual interactions. This approach is particularly promising when applied to Facebook groups.

The "Islam is Dangerous" Group

The "Islam is Dangerous" group is an "open" group on Facebook, which means that its shared posts and members are visible to every other Facebook user. At the time of writing, the group had 2339 members and was mainly dedicated to sharing information about atrocities, crimes, infractions or simply deviations from cultural standards by Muslims.

A first approach used Netvizz for extracting all friendship connections between all the members of the group. While it is difficult to imagine an "average" Facebook group, a first finding is constituted by what seems to be a relatively high network density of 0.019. An average degree of 39.7 is a second indicator that this is group hosts a tightly knit collective rather than a loosely associated group merely sharing information on a subject. Friendship patterns are, however, not evenly distributed. While 18.3% of the group members have no friendship connection with other members – a population attracted by the subject matter rather than through social contacts? – 37.2% have at least 20 connections and 14.8% 100 or more.

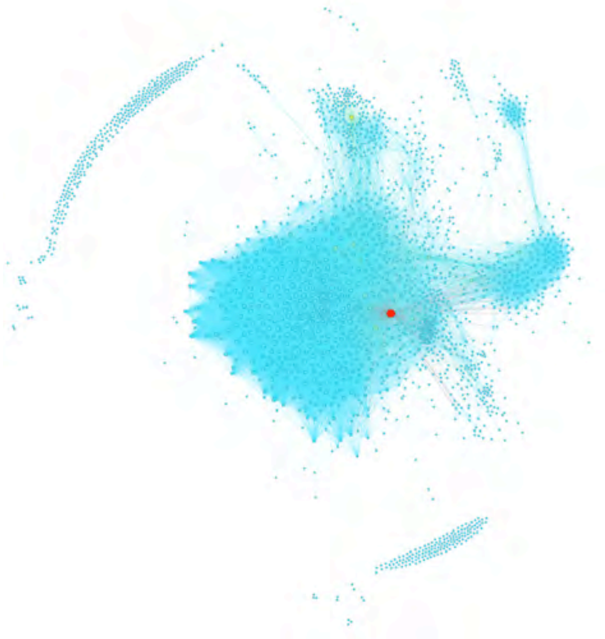


Figure 3. Friendship graph for the "Islam is Dangerous" group, colors represent betweenness centrality via a heat scale (blue => yellow => red).

While counting connections may be one way to identify leaders in a group, network analysis provides an extensive arsenal of techniques to analyze graphs in more specific ways. Figure 3 shows a spatialized visualization of the group (using gephi) and points to our ability to use advanced graph metrics to further analyzed the dataset by coloring nodes with a metric called *betweenness centrality*. This measure expresses a node's positioning in the larger topology of a graph and it can be very useful for detecting strategic positioning rather than popularity or social status. A person having high betweenness centrality is considered to be able to "influence the group by withholding or distorting information in transmission" [5] because he or she is located as a passage point between different sections of a network. While there are caveats to consider, betweenness centrality can be likened to Robert Putnam's concept of "bridging" social capital [10], which denotes the capacity to connect separate groups. In our case, this metric identifies the group administrator as the central *bridger*, which points to a group structure that, despite its high connectivity, is held together by a central figure.

The application of betweenness centrality can be seen as an example – a large number of techniques are now available to investigate structure, demarcate subgroups or qualify users in terms of their position in the network. Graph analysis software generally provides implementations of these metrics to researchers.

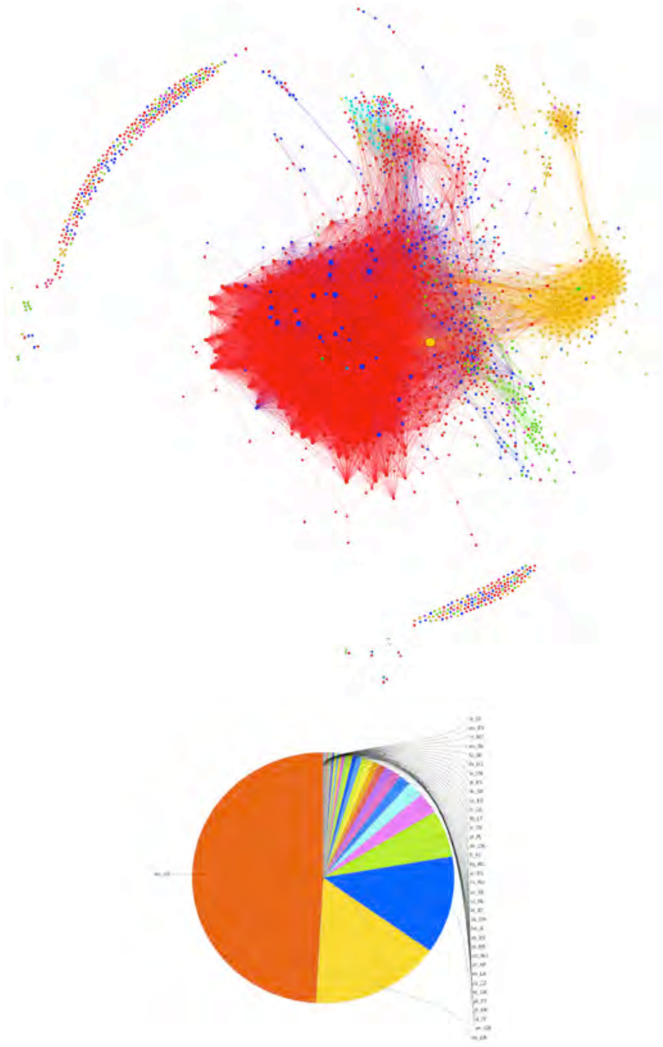


Figure 4. Friendship graph for the "Islam is Dangerous" group, colors represent "locale", i.e. the language of the Facebook interface for a given user.

Another example for types of analysis makes use of the users' interface language ("locale"), one of the few data points available for every Facebook member. Figure 4 shows the same network diagram as above, but uses locale to color nodes. We can see that there is a densely connected cluster of English speakers (both US and UK) that dominates the group, but smaller subcommunities, in particular a German one in yellow, can be identified as well. We can make the argument that this group, despite its high level of connectivity retains a degree of national coherence.

The "Educate children about the evils of islam" page

The second example quickly analyzes the Facebook page entitled "Educate children about the evils of Islam", which had been liked by 1586 users at the time of writing.

When extracting data from pages, Netvizz essentially operates by iterating over the last n (< 999) posts, collecting the posts themselves, as well as all of the users that like and comment on them. These data can be analyzed in various ways, either as bipartite network (Figure 5) or in more traditional form trough statistical analysis (Figures 6 and 7).

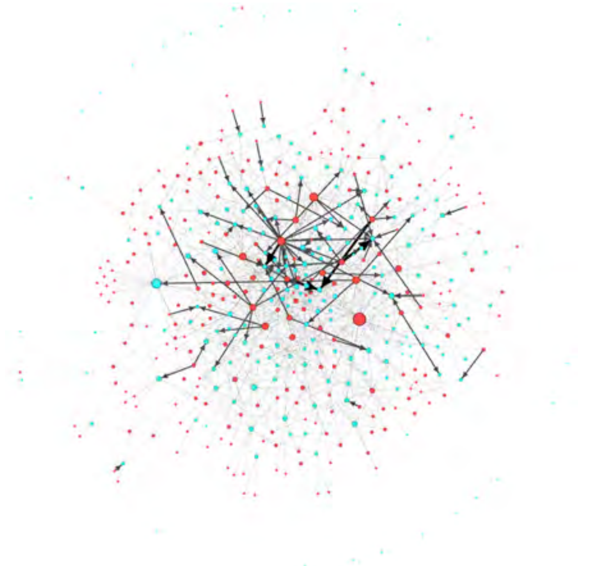


Figure 5. A network diagram showing the last 200 posts (turquoise), as well as the 253 users (red) liking and commenting them.

Network analysis maps interactions on a structural level and allows for the quick identification of particularly successful posts (in terms of engagement) and particularly active users. In this case, what emerges is a picture of a rather lively and intense conversational setting, with a core of loyal visitors that comment and react regularly.

Analyzing the posts over time (Figure 6), we can see that the 200 posts cover a period of less than four weeks, which indicates a high level of investment by the page owner, the only person allowed to post on the page.

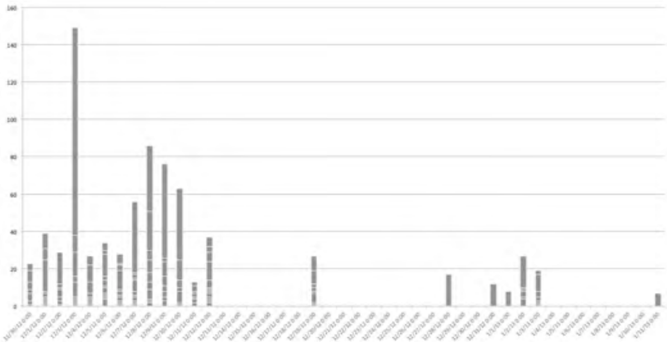


Figure 6. A stacked barchart showing the last 200 posts according to the days they were posted on; values indicate user engagement.

Because Facebook segments posts in content categories, we can also analyze content types, e.g. in relation to how particular types succeed in engaging users.

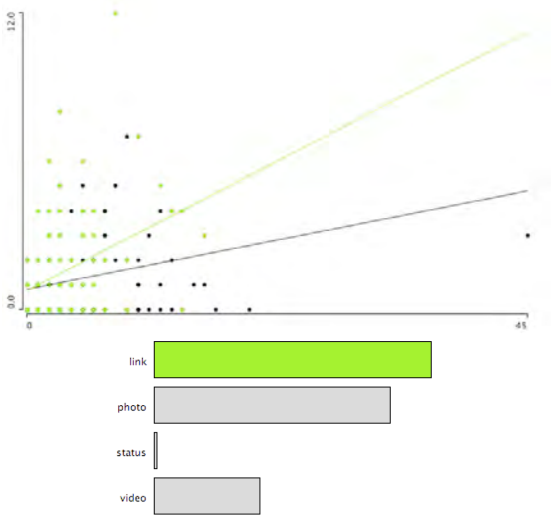


Figure 7. Visualization (using Mondrian) of the content types of the last 200 posts and how often they were liked (x-axis) and commented on (y-axis). Links are highlighted.

Figure 7 shows not only the distribution of content types over the last 200 posts (barchart), but also allows us to correlate these types to user activities. We can learn that links have a higher probability to receive comments, while photos are particularly likely to be liked.

These examples are mere illustrations of the analytical potential the in-depth data Facebook collects and Netvizz extracts. Many other types of analysis – from statistics to content analysis – are possible.

PRIVACY AND RESEARCH ETHICS CONSIDERATIONS

This final sections briefly sketches two aspects related to questions of privacy and research ethics, which would, however, merit a much more in-depth discussion that the space constraints allow.

The Facebook API as privacy challenge

Before discussing ethical considerations of data extraction on Facebook, it is useful to point out that part of the motivation for developing the Netvizz application was an exploration of the Facebook API itself, including the question how it governs access to data and what this means for users' capacity to limit or curate the way their data is accessible to others. This question is important because machine access needs to be treated differently than user interface access to data. While the latter is generally put to the front, the former allows for much more systematic forms of high speed and high volume data gleaning. Manual surveillance of activity is certainly possible, but I would argue that the largest part of user data collection by third parties on Facebook is performed via software that uses similar technological strategies as the Netvizz application. The application – and the knowledge gained by developing it – should therefore also be considered as an indicator of the types of information that other Facebook applications can get access to and certainly make extensive use of. While the fine-grained permission model holds the promise to limit third party access by asking users explicitly for permission, there is often no possibility for users to actually modulate which rights are granted: the application has to ask for detailed permissions for individual elements, but we can only acquiesce to all request or not use the platform. Access can be revoked *after* installation, but this means that applications can read that data at least once.

As Netvizz shows, a user granting rights to an application generally means that considerable access is given not only to her data, but also to *other* users' data. Application programming for research proposes is useful because of the analytical outcomes it produces or helps to produce, but it should also be considered as an investigation into the technological structures of platforms, which are as relevant to matters of privacy and beyond as they are understudied.

Research ethics

Social scientists have been confronted with the ethical dimension of empirical research well before the advent of the Internet. At no point have answers been easy or clear-cut. Recent debates amongst Internet researchers [20] have tended to put emphasis on the question of individual privacy. We should, however, note that there are significant cultural and political variations when it comes to arguing research ethics. Following Fuchs' critique [6] of the one-sided emphasis on a narrow definition of privacy, I would like to argue that research ethics navigate in a field defined by a number of tensions and competition between different ideals. Putting individuals' privacy on the top of the pyramid is a choice that can be traced to liberal sources of normative reasoning in particular, but we should not forget that these value sources are contingent and culturally colored. Competing ideals, such as the independence of research, larger social utility or the struggle against the encroaching of the private domain on publicness can equally be connected to established traditions in ethical reasoning.

It is clear that national traditions respond to these matters in different ways. While research ethics boards have become the norm in English-speaking countries, such an institutional governance of ethical decisions is hard to imagine in continental European countries such as France, where normative reasoning is concentrated both on the levels of the state and the individual, but only to a lesser degree on the layers in between. Similarly, the study of political extremism, and of the groups and individuals active in such movements, will not be framed in the same way in Germany and the United States, for obvious historical reasons.

What does that mean for Netvizz? Two decisions have been made: first, to anonymize all users for both groups and pages, simply because the number of accounts that can be collected this way is very large. For bigger pages, it is easy to quickly collect data for tens or even hundreds of thousands of user accounts. Second, Netvizz provides an option to anonymize accounts for personal networks. In this case, the complicated weighing of values and research ethics stays in the realm of the user/researcher and are only partially delegated to the programmer.

CONCLUSIONS

This paper has described the Netvizz application, a general-purpose data-extractor for different subsections of the Facebook platform. With a focus on questions relevant to media scholars, in particular, I have contextualized the application in a wider set of research concerns. With Facebook now counting over one billion active users, it is becoming urgent to develop and solidify research approaches to a service, largely constructed as a *walled garden*, that is part of an ongoing privatization of communication, both in terms of economics and accessibility. While there are important limits to what can be done without having to enter into a partnership with the company, the Netvizz application shows that certain parts of Facebook *are* amendable to empirical analysis, after all.

As Netvizz is continuously developed further, additional features will be added in the future. Providing more in-depth data on temporal aspects of user engagement with contents will certainly be one of the next steps.

ACKNOWLEDGMENTS

I would like to thank the attendants of the *Digital Methods Winter School 2013* for their useful comments, in particular Erik Borra and Jean-Christophe Plantin, as well as four anonymous reviewers.

REFERENCES

1. Agre, P.E. Surveillance and Capture: Two Models of Privacy. *The Information Society 10*, 2 (1994), 101-127.
2. Anscombe, F.J. Graphs in Statistical Analysis. *The American Statistician 27*, 1 (1973), 17-21.
3. Emerson, R.M. Social Exchange Theory. *Annual Review of Sociology 2*, (1976), 335-362.
4. Facebook Key Facts. http://newsroom.fb.com/Key-Facts.
5. Freeman, L.C. Centrality in Social Networks. Conceptual Clarification. *Social Networks 1*, 3 (1979), 215-239.
6. Fuchs, C. An Alternative View of Privacy on Facebook. *Information 2*, 4 (2011), 140-165.
7. Hacking, I. *Historical Ontology*. Harvard University Press, Cambridge, MA, USA, 2004.
8. Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A. and Christakis, N. Tastes, Ties, and Time: a New Social Network Dataset Using

Facebook.com. *Social Networks 30*, 4 (2008), 330-342.

9. Manovich, L. Trending: the Promises and the Challenges of Big Social Data. In Gold, M. *Debates in the Digital Humanities*. The University of Minnesota Press, Minneapolis, MN, USA, 2012, 460-475.

10. Putnam, R.D. *Bowling Alone: the Collapse and Revival of American Community*. Simon and Shuster, New York City, USA, 2000.

11. Quercia, D., Lambiotte, R., Kosinski, M., Stillwell, D., and Crowcroft, J. The Personality of Popular Facebook Users. In *Proc. CSCW 2012*, ACM Press (2012), 955-964.

12. Rogers, R. *The End of the Virtual*. Amsterdam University Press, Amsterdam, The Netherlands, 2009.

13. Rogers, R. Post-Democraphic Machines. In Dekker, A., Wolfsberger, A. *Walled Garden*. Virtual Platform, Amsterdam, The Netherlands, 2009, 29-39.

14. Schäfer, M.T. *Bastard Culture! How User Participation Transforms Cultural Production*. Amsterdam University Press, Amsterdam, The Netherlands, 2011.

15. Scott, J. Social Network Analysis. *Sociology 22*, 1 (1988), 109-127.

16. Simonite, T. What Facebook Knows. *MIT Technology Review*, June 13, 2012.

17. Ugander, J., Karrer, B., Backstrom, L. and Marlow, C. The Anatomy of the Facebook Social Graph. *eprint arXiv:1111.4503*, 2011.

18. Watts, D.J. The 'New' Science of Networks. *Annual Review of Sociology 30*, 1 (2004), 243-270.

19. Wilson, R.E., Gosling, S.D. and Graham, L.T. A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science 7*, 3 (2012), 203-220.

20. Zimmer, M. 'But the Data Is Already Public': on the Ethics of Research in Facebook. *Ethics and Information Technology 12*, 4 (2010), 313-325.

R. Rogers (2017). “Digital Methods for Cross-platform Analysis: Studying Co-linked, Inter-liked and Cross-hashtagged Content,” in Jean Burgess, Alice Marwick and Thomas Poell (eds.) *Sage Handbook of Social Media*. London: Sage, forthcoming.

Digital Methods for Cross-platform Analysis: Studying Co-linked, Inter-liked and Cross-hashtagged Content

Richard Rogers

Digital Methods after Social Media

By calling for a move from “so-called web 1.0 http or html approaches to 2.0 cross platform based methods,” Ganaele Langlois and Greg Elmer argue that to study the web these days requires new methods that step past the hyperlink as the preeminent digital object tying it all together (Langlois and Elmer, 2010:45). They issue a much larger invitation to rethink the web more generally as an object of study, recognising its increasing platformisation (Helmond, 2015). In the shift from an info-web (1.0) to a social web (2.0), recommendations are made by platform users rather by site webmasters (to use a throwback term). That is, recommendations, especially in the news feeds of platforms, follow from ‘friends’’ activity, such as ‘liking’ and ‘sharing’. The content recommendations thereby distinguish themselves epistemologically from those derived from site owners or webmasters' linking to another webpage for referencing or other purposes. Following Tim O’Reilly here the terms web 1.0 and web 2.0 have been used (or overused) to periodise not only the transition from the info-web to the social web, but also from the open web to the closed web or the walled gardens of platforms (O’Reilly, 2005; Dekker and Wolfsberger, 2009).

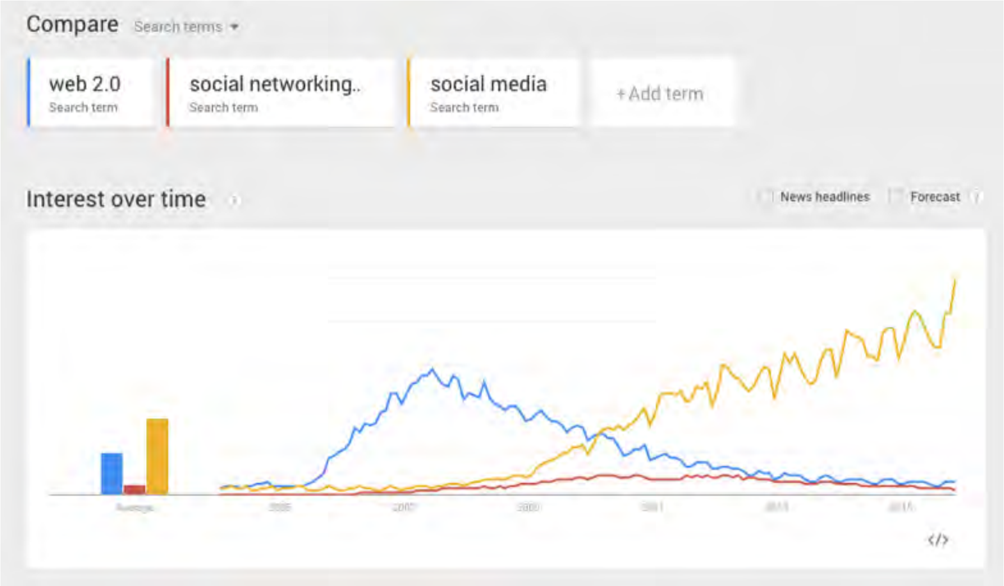


Figure One: Comparison of search volume for [web 2.0], [social networking sites] and [social media], according to Google Trends, 19 November 2015.

On the Web’s 25th anniversary in 2014, Tim Berners-Lee, who “slowly, but steadily” has come to be known as its inventor, called for its ‘re-decentralisation’, breaking down new media concentration and near monopolies online working as walled gardens without the heretofore open spirit (2014) (Agar, 2001: 371). The web’s ‘app-ification’ is analogous. Next to increased government Internet censorship, mass surveillance and punitive copyright laws, Berners-Lee lists ‘corporate walled gardens’ or social media platforms as grave concerns related to the very future of the web and its mobile counterpart.

Langlois and Elmer’s point, however, implies that one should not only periodise and critique the dominant phases of the web, but also do the same for its methods of study. There are those methods that rely on hyperlinks, and thereby in a sense still committed to an info-web, and those that have taken on board ‘likes’, ‘shares’ and other forms of valuation and currency (such as ‘comments’ and ‘liked comments’) on online platforms. Indeed, this analytical periodisation is reflected in the much broader study of value online, reflected in the rise of the ‘like economy’ over the ‘link economy’ (Gerlitz and Helmond, 2013). As a case in point, Google’s Web Search once valued links higher than other signals (Rieder, 2012). Through the rise of user clicks as source adjudication measure one could argue that Google Web Search, too, is valuing the social web over the document or semantic matching of the info-web (van Couvering, 2007). Metrification online, which starts with like counts and follower numbers and progresses towards Klout scores, similarly considers and makes rankings social. Thus the new analytics, both Google’s updated ones as well as Klout’s, are oriented to a web gone social.

Figure Two: One rendition of the Facebook Like button depicting a man’s hand, thumbs up, with a single-button barrel cuff. Originally the Like button was to be called the ‘Awesome’ button. See Bosworth, 2009. Image source: Wikipedia, 2015, https://upload.wikimedia.org/wikipedia/commons/1/13/Facebook_like_thumb.png.

The notion of web 2.0 (social web) brought with it as its apparent forerunner web 1.0 (info-web), but web 2.0 itself has been supplanted first by ‘social network(ing) sites’ and ‘platforms’ and later just by ‘social media’ (boyd and Ellison, 2007; Beer, 2008) (see figure one). The early distinction between social networking sites and social network sites, ushered in by boyd and Ellison, was normative as well as analytical. Social media users ought to have an interest to connect with others online other than for the purposes of ‘networking’, which would suggest a kind of neoliberal activity of making sure that even one’s social life (online) is productive. In a sense, the authors also anticipated the nuancing of social media into platform types, such as the ones for business (LinkedIn), family (Facebook) and professional doings (Twitter). Whether for networking or to connect with one’s existing network, the analytical call made by boyd and Ellison seemed to be directed to the study of profiles and friends (together with friending).

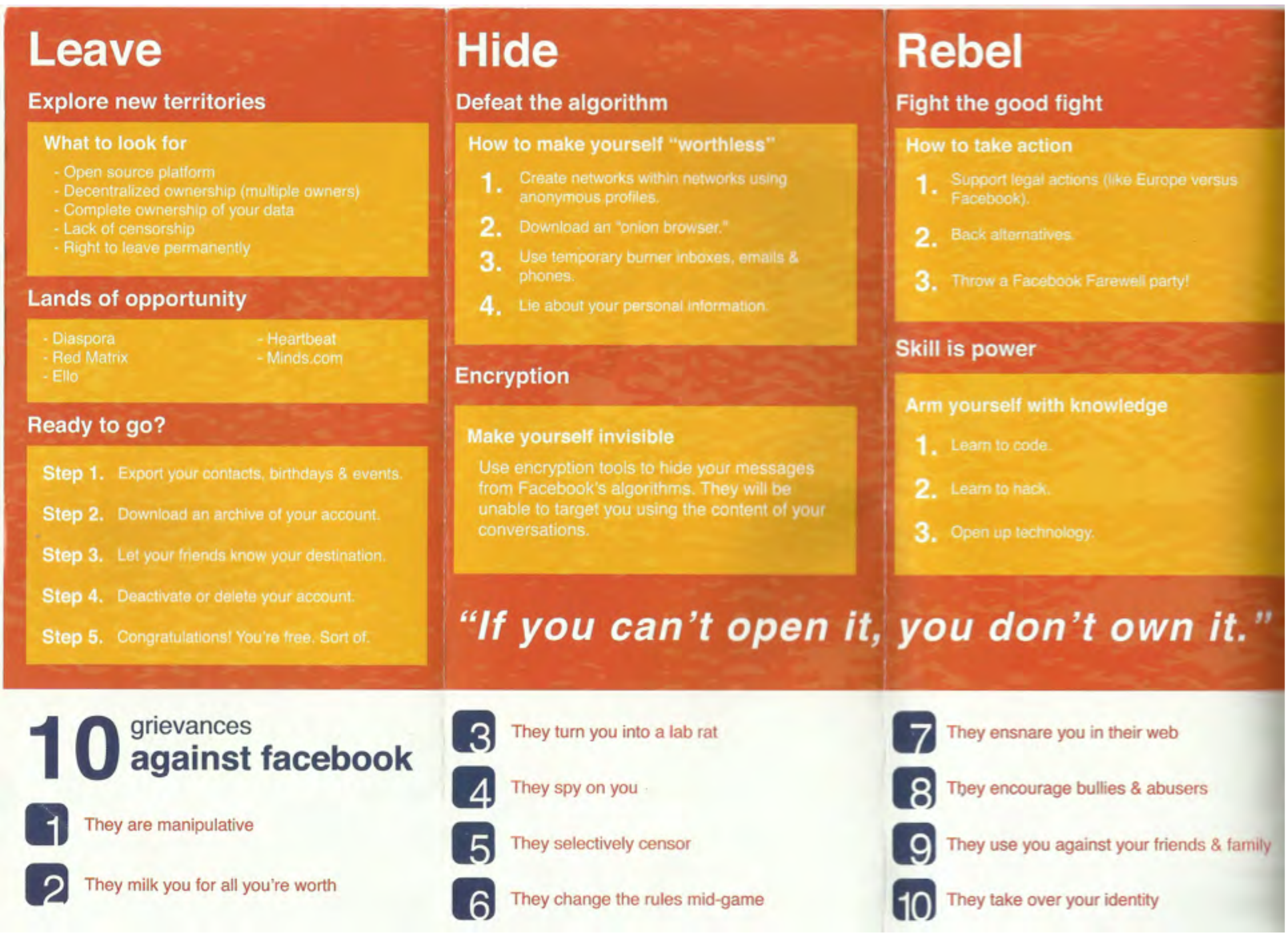
The purposive use of the term ‘platform’, as Tarleton Gillespie has pointed out, could be viewed as particularly enticing for users to populate an otherwise empty database, thereby generating value for the companies (2010). Platforms connote voice-giving infrastructure, where one can be express one’s viewpoints (political or otherwise), rise up, and make an online project of oneself. Polishing the profile, friending, uploading videos and photos, and liking, sharing and commenting become not only newly dominant forms of sociality but a kind of labour for a platform owned by others (Scholz, 2016). Cooperative, user-owned platforms would provide alternatives. Other critical calls for the analysis of Facebook have been made, certain of which have resulted in invitations to leave the platform, to liberate oneself or even to commit so-called Facebook suicide, which would allow you ‘to meet your real neighbours’, as suicidemachine.org's software project’s slogan has it. (See also figures three a and b.)

As web 2.0 has given way to social network(ing) sites, platforms and, finally, social media, ‘social media methods’ also have evolved. These methods initially relied on social network analysis (the study of interlinked friends) as well as profiles and the presentation of self. For example, Netvizz, the Facebook data extraction software, originally was considered a tool to map one’s own Facebook friend network (Rieder, 2013). The early digital methods work on social networking sites similarly studied friends and profiles. Dubbed ‘post-demographics’ this approach to studying profiles would consider preferences and tastes as a starting point of analysis as opposed to gender, age, education and such (Rogers, 2009). One case concentrates on the profiles of presidential candidates’ ‘friends’, and considers whether they listed as interests the same or quite distinctive television shows, movies, heroes, and books. As the question read in an analysis of MySpace, do Barack Obama’s friends and John McCain’s friends share the same interests, or is there a distinctive politics to media taste and consumption? Are we able to study productively (or reanimate) the culture wars via aggregations of social media profiles? (For the most part they did not share tastes and thus TV shows and the other preferences could be considered to have politics of consumption.) In the case of Netvizz friend-network mapping as well as post-demographics these methods could be called social media method 1.0.

Figure Three A: Facebook liberation army flyer. Source: fla.waag.org.

More recently attention in social media method has been directed towards events, disasters, elections and revolutions, first through the so-called ‘Twitter revolution’ surrounding the Iran election crisis (2009) and later the Arab Spring (2011-2012). Instead of user profiles and networks or networking, the starting point would be a tweet collection curated through one or more hashtags such as #iranelection (perhaps together with queried keywords), or a well-liked Facebook page, such as We are all Khaled Said (Gaffney, 2010; Lotan et al., 2011; Rieder et al., 2015).

In any discussion of a second phase of social media analysis (from the study of profiles and friends to that of events, disasters, elections, revolutions and social causes), it should be pointed out that many of the more recent methods to analyse platforms rest upon and also derive from the individual APIs Twitter, Facebook, Instagram, YouTube and others have to offer. As data are increasingly offered and delivered by polling one API, and no longer screen-scraped or crawled from multiple websites (as in the days of the info-web), most work is a study of a page or multiple pages (and groups) on Facebook, or one concerning tweets containing one or more hashtags or keywords on Twitter. In social media method, in other words, 'single-platform studies' have become the norm.



Leave

Explore new territories

What to look for

- Open source platform
- Decentralized ownership (multiple owners)
- Complete ownership of your data
- Lack of censorship
- Right to leave permanently

Lands of opportunity

- Diaspora
- Red Matrix
- Ello
- Heartbeat
- Minds.com

Ready to go?

Step 1. Export your contacts, birthdays & events.

Step 2. Download an archive of your account.

Step 3. Let your friends know your destination.

Step 4. Deactivate or delete your account.

Step 5. Congratulations! You're free. Sort of.

Hide

Defeat the algorithm

How to make yourself "worthless"

1. Create networks within networks using anonymous profiles.
2. Download an "onion browser."
3. Use temporary burner inboxes, emails & phones.
4. Lie about your personal information.

Encryption

Make yourself invisible

Use encryption tools to hide your messages from Facebook's algorithms. They will be unable to target you using the content of your conversations.

Rebel

Fight the good fight

How to take action

1. Support legal actions (like Europe versus Facebook).
2. Back alternatives.
3. Throw a Facebook Farewell party!

Skill is power

Arm yourself with knowledge

1. Learn to code.
2. Learn to hack.
3. Open up technology.

"If you can't open it, you don't own it."

10 grievances against facebook

1 They are manipulative

2 They milk you for all you're worth

3 They turn you into a lab rat

4 They spy on you

5 They selectively censor

6 They change the rules mid-game

7 They ensnare you in their web

8 They encourage bullies & abusers

9 They use you against your friends & family

10 They take over your identity

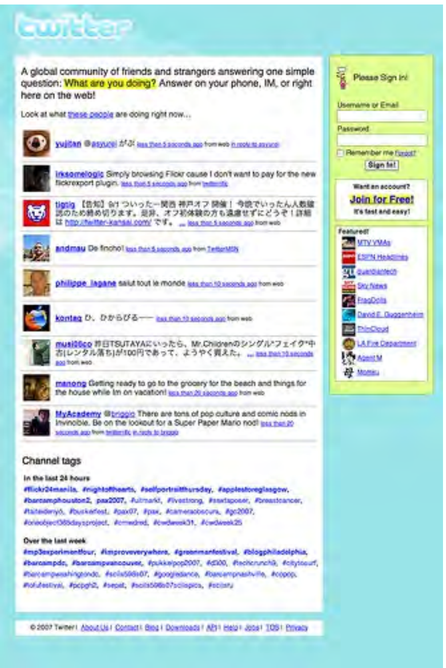
Figure Three B: Facebook liberation army flyer, with so-called directives, instructions and grievances. Source: fla.waag.org.

If there were a significant turning point towards single-platform studies with the API, together with the API ultimately steering the work that can be undertaken, it may have been the critique of a social network study of Facebook data. It concerned a set of presumably anonymised users from a so-called renowned university in the northeast of the United States (Lewis et al., 2008; Zimmer, 2010). Not so unlike the effects of the release of AOL user search histories in 2006, its publishing prompted detective work to uncover the identities of the users, who turned out to be Harvard College students from the graduating class of 2009 (Zimmer, 2008). Michael Zimmer, both in the detective work as well as in the reflection upon the way forward for social media method, entitled his critique, ‘But the data is already public’, echoing one of the remarks of an author of the study. In giving rise to a sharper focus on ethics in web studies more generally, coinciding with a decline in scraping, Zimmer argued that in the Harvard study users’ so-called contextual privacy was violated, for not only did they not give informed content but they did not expect their publicly available data to be stored in a researcher’s database and matched with their student housing data for even greater analytical scrutiny of their ‘ties and tastes’, the subject of the study (Nissenbaum, 2009). The actual data collection is described by the researchers as ‘downloading’ the profile and friend network data directly from Facebook, prior to the release of Facebook API 1.0 in 2010. In other words the data were obtained or scraped in some non-API manner, albeit with permission from Facebook as well as Harvard for the project funded by the National Science Foundation and approved by the university’s ethics review board. Ultimately in the evolution of its API to version 2.0 (in 2014), Facebook would remove permissions to access friends’ data such as ties and tastes (i.e., friends and likes, together with profiles), thereby making (sociometric) social network analysis like the one performed in the Harvard study improbable, including even those of one’s own network with all friends’ privacy settings adhered to (Facebook, 2016). ‘Internal’ studies still may be performed, which Facebook data scientists also took advantage of with their ‘emotional contagion’ experiment (Kramer et al., 2014). The data science study (of some 700,000 users with a corpus of 3 million posts) analysed the risks associated with the Facebook news feed. Is user exposure to positive or negative posts psychologically risky (Meyer, 2015)? The study found that negative posts run the risk of ‘emotional contagion’. In order to make the findings, Facebook selectively removed negative posts from users’ news feeds. The ethics of the study were similarly questioned, for the users were unaware (and not informed) that their news feeds were being altered and their moods measured, however seemingly impractical and obtrusive it would be to gain such permission (Puschmann and Bozdag, 2014). Among the ethical issues raised concerned whether researchers can rely on the terms of service as cover for the otherwise lack of informed consent. Are users agreeing to being analysed for more than improvement of the site and services, as is usually stated? To the letter, they are not.

It is worthwhile to recall from the AOL case that the senior citizen from Georgia told the *New York Times* that she never imagined that her search engine queries would be made public, or would have to explain to anyone that her information-seeking about medical conditions she undertook for her friends, too. In joining a lawsuit brought against AOL at the Federal Trade Commission, the Electronic Frontier Foundation published highly personal and salacious query histories (from unnamed individuals); another user’s was made into the mini-documentary, “I Love Alaska: The heartbreaking search history of AOL user #711391”, by the Dutch artists and filmmakers, Lernert Engelberts and Sander Plug, who were asked subsequently by the broadcasting company to seek out the identity of the woman, now intimately known (2009). (Ultimately they did not.) Neither of the social media studies (Harvard’s graduating class of 2009 and Facebook’s emotional contagion) appears to have led to the subjects being identified and in some way harmed through outing. It is also not straightforward to claim that informed consent would have been enough to preclude harm, given that the users may be unable to foresee the potential hazards of participation (van de Poel, 2009).

Hashtag and (Liked) Page Studies

With the decline of scraping and the rise of issues surrounding human subject research in social media, the API-led studies (on events, disasters, elections, revolutions and social causes) rely increasingly on such content-organising elements as the hashtag (for Twitter) and the (liked) page (for Facebook). Each is taken in turn, so as eventually to discuss with which limitations one may study them concurrently across platforms.



A global community of friends and strangers answering one simple question: What are you doing? Answer on your phone, IM, or right here on the web!

Look at what these people are doing right now...

yujitan @asyuraj がふ less than 5 seconds ago from web in reply to asyuraj

irksomelogic Simply browsing Flickr cause I don't want to pay for the new flickrexport plugin. less than 5 seconds ago from twitterrific

tigtig 【告知】9/1 ついった一関西 神戸オフ 開催！ 今晩でいったん人数確認のため締め切ります。是非、オフ初体験の方も遠慮せずにどうぞ！詳細は http://twitter-kansai.com/ です。 ... less than 5 seconds ago from web

andmau De fincho! less than 5 seconds ago from TwitterMSN

philippe_lagane salut tout le monde less than 10 seconds ago from web

kontaq ひ、ひからびる―― less than 10 seconds ago from web

musi26co 昨日TSUTAYAにいったら、Mr.Childrenのシングル"フェイク"中古(レンタル落ち)が100円であって、ようやく買えた。 ... less than 10 seconds ago from web

manong Getting ready to go to the grocery for the beach and things for the house while Im on vacation! less than 20 seconds ago from web

MyAcademy @briggig There are tons of pop culture and comic nods in Invincible. Be on the lookout for a Super Paper Mario nod! less than 20 seconds ago from twitterrific in reply to briggig

Channel tags

In the last 24 hours

#flickr24manila, #nightofhearts, #selfportraitthursday, #applestoreglasgow, #barcamphouston2, pax2007, #ultmarkt, #livestrong, #saetaposer, #breastcancer, #taitaidenyo, #buskerfest, #pax07, #pax, #cameraobscura, #gc2007, #oneobject365daysproject, #cmedred, #cwdweek31, #cwdweek25

Over the last week

#mp3experimentfour, #improveverywhere, #greenmanfestival, #blogphiladelphia, #barcampdc, #barcampvancouver, #pukkelpop2007, #d300, #techcrunch8, #citytosurf, #barcampwashingtondc, #scils598s07, #googledance, #barcampnashville, #coppop, #lolufestival, #pcpgh2, #sepat, #scils598s07sclispics, #scilsfu

Please Sign In!

Username or Email

Password

□ Remember me Forgot?

Sign In!

Want an account?

Join for Free!

It's fast and easy!

Featured!

MTV VMAs

ESPN Headlines

quantumtech

Sky News

FragDolls

David E. Guggenheim

ThinCloud

LA Fire Department

Agent M

Momeu

© 2007 Twitter | About Us | Contact | Blog | Downloads | API | Help | Jobs | TOS | Privacy

Figure Four. Proto-history of Twitter's Trending Topics. Chris Messina's proposal for Twitter 'Channel Tags,' ranked by most active, August 2007.

The Twitter hashtag, put forward by Chris Messina in 2007, originally was conceived as a means to set up 'channel tags,' borrowing from similar practices in Internet Relay Chat (IRC) (see figure 4). The proposal was to organise "group-like activity" on Twitter that would be "folksonomic," meaning user-organised rather than an editorial or taxonomic practice by the company or its syndicated partners as in Snapchat's 'Stories' (Messina, 2007). Messina also proposed to provide a ranked list of the channel tags by activity, i.e., most active ones in the past twenty-four hours, showing on the interface where the activity is. It is an feature similar to trending topics which Jack Dorsey, co-founder of Twitter, a year later described as "what the world considers important in this moment" (Dorsey, 2008). With hashtags and trending topics, Twitter not only gained new functionality but became a rather novel object of study for what could be termed both on-the-ground but also 'remote event analysis.' As such it thus distinguishes itself from Dorsey's original Twitter, created to provide what he called "personal immediacy — seeing what's happening in my world right now" (Dorsey, 2008). Dorsey himself acknowledged the shift away from his more intimate Twitter in the interviews he gave for the *Los Angeles Times* after his temporary ouster as CEO, saying Twitter thrives on "natural disasters, man-made disasters, events, conferences, presidential elections" (Sarno, 2009). In the event, the study of Twitter as a space for ambient friend-following yielded, at least for a large share of Twitter studies, to that of event-following, which is another way of distinguishing between social media method 1.0 and 2.0.

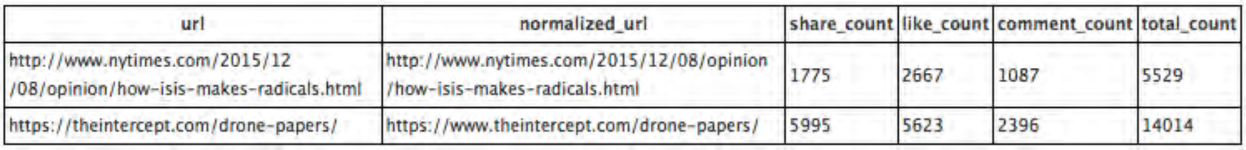
Not so unlike Google Trends that list the year's most sought key words (with a geographical distribution), Twitter's initial cumulative list of the year's trending topics, published in 2009, provides a rationale for the attention granted to the study of the single hashtag for events. In the announcement made by the Twitter data scientist, Abdur Chowdhury (who incidentally was head of AOL Research when the search history data were released), one notes how serious content began to take a prominent place in a service once known primarily for its banality (Rogers, 2013). In 2009 "Twitter users found the Iranian elections the most engaging topic of the year. The terms #iranelection, Iran and Tehran were all in the top-21 of Trending Topics, and #iranelection finished in a close second behind the regular weekly favorite #musicmonday" (Chowdhury, 2009). Some years later the universal list of trending topics became personalised according to whom one follows and one's geographical coordinates, however much one may change one's location and personalise trending topics exclusively by new location. In some sense the change from universal to personalised results (like Google Web Search's similar move in December, 2009, which Eli Pariser relies upon for his notion the 'filter bubble') made trends more unassailable, for no longer could one call into question why a particular hashtag (like #occupywallstreet) was not trending when it perhaps should have been (Gillespie, 2012). Trending topics are in a sense now co-authored by the Twitter user, making them less compelling to study at least as a cultural barometer. (The exception is trending topics that are location-based only.)

Whilst the single hashtag, or more likely a combination of hashtags and keywords, remain a prominent starting point for making tweet collections to study such events, disasters, elections, revolutions and social causes, together with subcultures, movements, stock prices, celebrity awards and cities, researchers have expanded widely their repertoire for assembling them, first through techniques of capturing follower, reply and mention networks, and subsequently using the 1% random sample (made available by Twitter), geotagged tweets and the Twitter ID numberspace in combination with timezones to identify national twitter spheres (Gerlitz and Rieder, 2013; Crampton et al. 2013; Bruns et al. 2015).

Network analysis remains a preferred analytical technique, and as such it endures in the transition to method 2.0, but one somewhat novel strand of work worthy of mention here concerns Twitter content studies, discussed by way of a brief analytical tool description.

The Twitter Capture and Analysis Tool (TCAT) by the Digital Methods Initiative one installs on one's own server to capture tweets for analysis. Individual researchers thereby make individual tweet collections, instead of having one or more larger databases that are collaboratory-like repositories. Such archival fragmentation could not be avoided, because Twitter, once rather open, changed its terms of service upon becoming a publicly traded company, no longer allowed the sharing of tweet collections (Puschmann and Burgess, 2013). Thus researchers must make their own tweet collections. The TCAT tool, installed on a server (with GitHub instructions), enables tweet collection-making (gathered from the search API) and provides a battery of network analyses: social graph by mentions, social graph by in_reply to status_id, co-hashtag, bipartite hashtag-user, bipartite hashtag-mention, bipartite hashtag-URL and bipartite hashtag-host. There are also modules, however, that direct attention towards forms of content analysis that are 'quanti-quali' and referred to as 'networked content analysis' (Niederer, 2016). By quanti-quali is meant that a quantitative, winnowing analysis (not so unlike sampling) is performed so as to enable not only a 'computational hermeneutics' but also a thicker description (Mohr et al., 2015). Quanti-quali is preferred over the more usual quali-quanti moniker, owing to the order of the methodological steps (Venturini et al., 2015). Departing from a collection of 600,000 tweets gathered through a single hashtag, an example of such an approach is the #iranelection RT project, which sought to turn Twitter into a story-telling machine of events on the ground and in social media by ordering the top three retweeted tweets per day, and placing them in chronological order, as opposed to the reverse chronological order of Twitter (Rogers et al., 2009). #iranelection RT relied on manual retweeting (where the user types RT in the tweet), whereas the TCAT module outputs, chronologically, 'identical tweet frequency', or narrowly defined retweets. Other forms of quanti-quali content analysis with a tweet collection are hashtag as well as URL frequency list-making so as to study hierarchies of concern and most referred to content. It is the starting point for a form of content analysis that treats a hashtag as (for example) an embedded social cause or movement (#blacklivesmatter) and URLs a webpage such as a news story or YouTube video. The (often fleeting) 'hashtag publics' mobilise around a social cause not only phatically (and affectively) but also with content (Bruns and Burgess, 2011; Papacharissi, 2015). Networked content analysis considers how and to what substantive ends the network filters stories, mobilises particular media formats over others and circulates urgency (geographically), attracting bursty or sustained attention that may be measured. Techniques of studying social causes using hashtags in Twitter as well as Instagram are discussed below, including how to consider whether to downplay or embrace medium effects.

Whilst since June 2013 Facebook has included hashtags as proposed means of organising ‘public conversations,’ the straightforward ‘cross-platform analysis’ of Twitter and Facebook using the same hashtags is likely fraught. The study of Facebook ‘content’ relies far more on other activities as liking, sharing and commenting, which is known as studying ‘most engaged with content’ (and is available in the Netvizz data outputs) (see figure five). For cross-platform work, the co-appearances of URLs (aka co-links) amplified perhaps by ‘likes’ (Facebook’s as well as Twitter’s new favourites) may yield far more material for comparative resonance analysis.



url	normalized_url	share_count	like_count	comment_count	total_count
http://www.nytimes.com/2015/12 /08/opinion/how-isis-makes-radicals.html	http://www.nytimes.com/2015/12/08/opinion /how-isis-makes-radicals.html	1775	2667	1087	5529
https://theintercept.com/drone-papers/	https://www.theintercept.com/drone-papers/	5995	5623	2396	14014

Figure Five: Netvizz output showing the resonance of two URLs on Facebook.

From the beginning Facebook (unlike Friendster and MySpace before it) positioned itself as a social network site that would reflect one’s own proper circle of friends and acquaintances, thereby challenging the idea that online friends should be considered ‘friends’ with quotation marks and thereby a problematic category worthy of special ‘virtual’ study. In a sense such a friend designation could be interpreted as another mid-2000 marker of the end of cyberspace. Together with the demise of serendipitous (and aimless) surfing, the rise of national jurisdictions legislating (and censoring) the Internet and the reassertion of local language (and local advertising) as organising principles of browsing, Facebook also re-ordered the web, doing away with cyberspace in at least two senses. As AOL once did with its portal, Facebook sought to attract and keep users by making the web ‘safe,’ first as a U.S. college website offering registration only to on-campus users with an .edu email address, and then later as it expanded beyond the colleges by ID-ing users or otherwise thwarting practices of anonymisation (Stutzman et al. 2013). Facebook’s was an effort to prevent fakesters, and thus distinguish itself from the other online platforms with their lurkers, stalkers as well as publicised cases of sex offenders masquerading as youngsters. Facebook’s web was also clean, swept of visual clutter. In contrast to MySpace, it did not offer customisation, skinning or pimping, so one’s profile picture and the friend thumbnails would be set in a streamlined, blue interface without starry nights, unicorns and double rainbows surrounding the posts.

Facebook’s safe and de-cluttered web brought a series of ‘cyberspace’ research practices down to earth as well, cleaning up or at least making seem uncouth such practices as scraping websites for data (Marres and Weltevrede, 2013). For one, scraping social network sites for data became a (privacy and proprietary) concern and also a practice actively blocked by Facebook. Data would be served on Facebook’s terms through its API (as mentioned above), and the politics and practices of APIs (more generally) would become objects of study (Buchner, 2013). In this case terms-of-service-abiding, non-scraping data extraction tools (such as Netvizz) would reside on Facebook itself, and require vetting and approval by the company. Be it through the developers’ gateway or a tool on Facebook, one would log in, and the data available would respect one’s own as well as the other users’ privacy settings, eventually putting paid to the open-ended opportunities social network sites were thought to provide to social network research. With the API as point of access, Facebook as an object of study has undergone a transition from the primacy of the profile and friends’ networks (‘tastes and ties’) to that of the page or group, and with it from the presentation of self to social causes (which I’m using as a shorthand for events, disasters, elections, revolutions and so forth). In a sense the company's acquisition, Instagram, could be said to have supplanted Facebook as the preferred object of study of the self through its ambassadorship of selfie culture, however much its initiator would like the company to take the route of Twitter, at once debanalising and becoming a news and event-following medium (Senft and Baym, 2015; Goel, 2015).

If, with the API, Facebook analysis is steered towards the pages of social causes, 'liking' is no longer considered as frivolous. As a case in point liking a page with photos of brutal acts of violence requires the like button to be re-appropriated, as Amnesty International (and other advocacy organisations) are wont to do by asking one not to take liking lightly (or communicate only phatically) but to see liking as an act of solidarity with a cause or support for a campaign. Whilst it has been dismissed as a form of slacktivism (which requires little or no effort and has little or no effect), liking as a form of engagement has been studied more extensively, with scholars attributing to button clicking on Facebook distinctive forms of liking causes: "(1) socially responsible liking, (2) emotional liking, (3) informational liking, (3) social performative liking, (5) low-cost liking and 6) routine liking" (Brandtzaeg and Haugstveit, 2014: 258). In the event, low-cost liking would be especially slacktivist, though all forms of liking in the list also could be construed as a form of attention-granting with scant impact, as was once said of the 'CNN effect' when all the world's proverbial eyes are watching — but not acting (Robinson, 2002). The question of whether liking as a form of engagement substitutes for other forms, however, has been challenged, for social media activism, it is argued, aids in accumulating action and action potential (Christensen, 2011). It is also where the people are (online).

From single platform to cross-platform studies

Social movement, collective action and more recently 'connective action' researchers in particular have long called for multiple platform, and multi-media, analysis (to use an older term). In an extensive study based on interviews, Sasha Costanza-Chock, for one, has deemed the immigrant rights movement in the United States a form of 'transmedia organising' (2014). The cross-platform approach is a deliberate strategy, and each platform is approached and utilised separately for its own qualities and opportunities. Here one may recall the distinction made by Henry Jenkins between cross-media (same story for all platforms) and transmedia (the story unfolds across platforms) (2006). Thus social media, when used as a "collapsed category", masks significant differences in "affordances" (Costanza-Chock, 2014: 61-66). (I return to a similar problem concerning collapsed digital objects such as hashtags or likes across platforms with different user cultures.) If we are to follow Jenkins, as well as Costanza-Chock, a discussion of cross-platform analysis would be more aptly described as trans-platform analysis.

Other researchers (studying social causes on platforms) also have called for 'uncollapsing' social media. Lance Bennett and Alexandra Segerberg, who coined the notion of 'connective action' as a counter-point to collective action, argue that to understand the forces behind social change one should study *those multiple platforms* that allow for 'personalized public engagement', instead of choosing one platform and its API in advance of the analysis (2012). It is, in other words, an implicit critique of the single-platform studies (as collapsed social media studies) that rely solely on Twitter for one issue (e.g., Fukushima in Japan) or Facebook for another (e.g., rise of right-wing populism), when one could have ample cause to study them across media. It is not only the silo-ing of APIs that prompts single-platform studies; as pointed out, the question of the comparability of the 'same' objects across platforms (likes, hashtags) is at issue.

One of Bennett and Segerberg's preferred tools is the Issuecrawler, developed at the Digital Methods Initiative, which could be described as web 1.0 analytical software, relying on the info-web's link and performing hyperlink analysis. For multiple-platform (and transmedia) analysis à la Bennett and Segerberg it could be employed as an exploratory instrument at the outset of a study of a cause (on the web) in order to ascertain which websites (including blogs) and platforms are the focus of attention. In other words, hyperlink analysis could be construed as a web 1.0 methodological starting point for multi-platform analysis. As described below, other ‘interlinkings’ (broadly conceived) may be studied, such as co-linked and inter-liked content.

Figure Six: A graphical reaction the transformation of Twitter’s favorite button (star) to a like button, announced by Akarshan Kumar, “Hearts on Twitter (and Vine),” *Twitter blog*, 3 November 2015, https://blog.twitter.com/2015/hearts-on-twitter.

Platform cultures of use

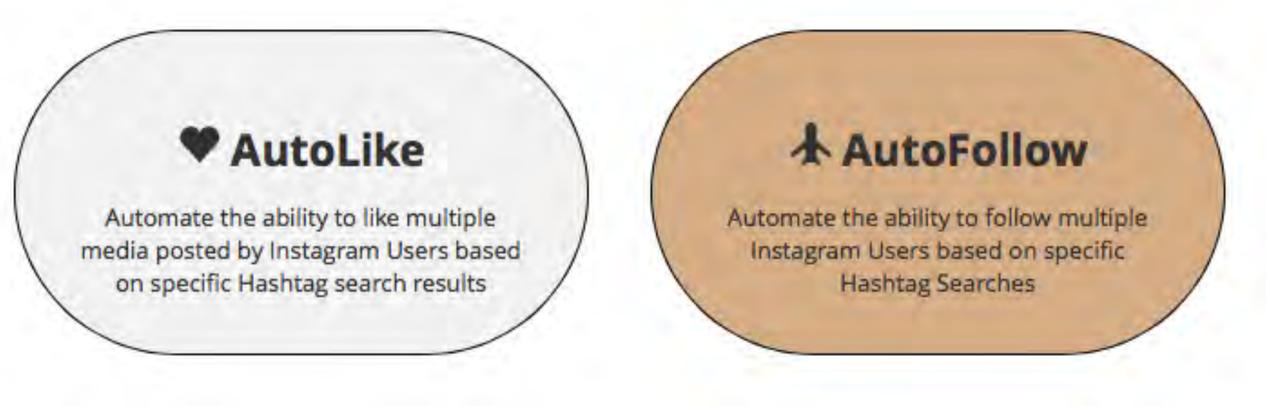
The purpose of the exercise here is to develop cross-platform methods, or digital methods for cross-platform studies, where one learns from medium methods and repurposes them for social and cultural research. It begins with a sensitivity to distinctive user cultures and subcultures, whereby hashtags and likes, used to organise and boost content (among other reasons), should not necessarily be treated as if they are employed equivalently across all platforms, even when present. For example, Instagram has inflated hashtag use compared to Twitter’s, allowing up to thirty tags (and far more characters per photo caption post than Twitter grants for a tweet). That is, users may copy and paste copious quantities of hashtags in Instagram posts (see table one). Twitter recommends that one “[does not] #spam #with #hashtags. Don't over-tag a single Tweet. (Best practices recommend using no more than 2 hashtags per Tweet.)” (Twitter, 2016). Whilst present, hashtags are under-utilised on Facebook.

Table one: Sample of suggested tags to copy and paste as caption for an Instagram photo, in order to garner more likes and followers, as is claimed. Category of tags: ‘most popular’. Source: http://tagsforlikes.com, 4 December 2015.

```
#love #amazing #smile #follow4follow #like4like #look #instalike
#igers #picoftheday #food #instadaily #instafollow #followme #girl
#iphoneonly #instagood #bestoftheday #instacool #instago
#all_shots #follow #webstagram #colorful #style #swag
```

A series of questions arises concerning the ‘cross’ in cross-platform analysis. First, across which platforms are ‘hashtags’ worthy of study (Twitter, Instagram, Tumblr), which ones ‘likes’ (Facebook, Instagram, YouTube, Twitter, Pinterest), which ones retweets or repins (Twitter, Pinterest), which one ‘@mentions’ (Twitter), which ones ‘links’ including shortened URLs (not Instagram) and so forth (see Table two)? Here the point is that platforms have on offer the same or similar digital objects for clicking or typing (like buttons, hashtags), but one should not necessarily collapse them by treating them equally across platforms. More specifically, if one were to perform cross-platform analysis of the same hashtags across multiple platforms, how would one build into the method the difference in hashtag use in Twitter and Instagram? Because of hashtag proliferation on Instagram, does one devalue or otherwise correct for hashtag abundance on the one platform whilst valuing it steadily on another? One could strive to identify cases of copy-and-pasting hashtag strings, and downplay their value, certainly if posts are being ‘stuffed’ with hashtags.

Indeed, secondly, certain platforms (and perhaps more so certain topics such as large media events on most any platform) may have user cultures and automation activity that routinely befoul posts as well as activity measures. Hashtag hijacking is a case in point, especially when one is studying an event or a social issue and encounters unrelated hashtags purposively inserted to attract attention and traffic. Hashtag junk may distract, at least the researcher.



♥ AutoLike

Automate the ability to like multiple media posted by Instagram Users based on specific Hashtag search results

AutoFollow

Automate the ability to follow multiple Instagram Users based on specific Hashtag Searches

Figure Seven: Features of iFollowandLike, the Instagram bot, that takes the work out of liking and following through automation. Source: Screenshot from iFollowandLike.com, 4 December 2015.

Thirdly, whilst a more complex topic, bots and the activity traces they leave behind are often similarly considered worth flagging during the analysis. From a digital forensics point of view bots that like and follow may have specific signatures (e.g., they do not tend to be followed, or to be liked, thus leaving star shapes) (see figure seven). For the purposes of this discussion, they may inflate activity in causes and such inflation may be considered artificial (though of course there are bots created for events and issues, too, and their activities are thereby purposive). Thus manipulation as well as artificiality are additional (intriguing) complications in both single-platform and cross-platform analysis.

Fourthly, platforms have ‘device cultures’ as well. These are how users may interact with the interface given how data collected on the users feed back (through the algorithmic system) on what users see.[1] That is, all platforms filter posts, showing particular content and letting other content slide, so to speak (Eslami, 2015). Users thereby cannot ‘like’ all content equally. That which is liked may tend to be liked more often, and thus there may be power law and long tail effects that differ per platform. But we may not know how preferred posting affects activity measures. APIs will return like and share counts (for example) per post, but they do not let us know the extent to which all the content has been equally visible to those who would be able to like, share, comment and so forth. And filtering styles and thus visibility effects differ per platform.

Above a series of questions has been posed concerning the limitations of comparing evaluations of content, recommended with the same type of button on different platforms, given that the platforms may have different user, spamming, bot and device cultures. How to nevertheless undertake cross-platform analysis? When studying recommendations and

[1] ’Device culture’ studies would inquire into the chain of interactions between user and platform that results in data collected and system-analysed so that ultimately content is recommended recursively back to the user (Rogers et al., 2013; Weltevrede, 2016).

the content that rises, metrically, to the top of the platforms, it may be instructive to begin by examining briefly which digital objects are available in each of the platforms (as above and in Table two) and subsequently enquire into how dominant devices (or in this case metrics such as Klout) handle these objects. Subsequently, it is asked, how to repurpose the metrics?

Cross-platform analysis: Co-linked, inter-liked and cross-hashtagged content
Klout, as the term indicates, measures a user's 'clout', slang for influence, largely from data culled online, where the user is not only an individual but can be a magazine, institution, professional sports team, etc. Klout scores are measured on the basis of activity on Twitter, Facebook, YouTube, Google+, LinkedIn, Instagram, and Foursquare (Rao et al., 2015). It is an influence measure that takes into account particular appearance signals across the seven platforms (e.g., mentions on Twitter), and those mentions by highly influential useraccounts grant more influence or clout to the user in question (see figure seven). It also grounds (and augments) the online appearance measures with "offline factors" that take into account a user's "real world influence" from Wikipedia as well as resonance in news articles (Rao et al., 2015: 3). Job titles, years of experience and similar from LinkedIn are also factored in. It is also a computationally intensive, big data undertaking.

If one were to learn from Klout for social research, one manner would be to shift the focus from power (measures of increases or decreases in one's influence) to matters of concern (increases or decreases in attention, including that from significant others) — be these to events, disasters, elections, revolutions, social causes and so forth. The shift in focus would be in keeping with how social media is often currently studied, as discussed above. That is, one could apply Klout's general procedure for counting user appearances, and ask, which causes are collectively significant across social media platforms, and which (key) actors, organisations and other users are linked to them, thereby granting them attention? Just as importantly, the attention granted to a cause by key actors, organisations and users may be neither undivided nor sustained. Such an observation would invite inquiries into partial attention as well as attention span, which together could begin to form a means to study engagement in social media.

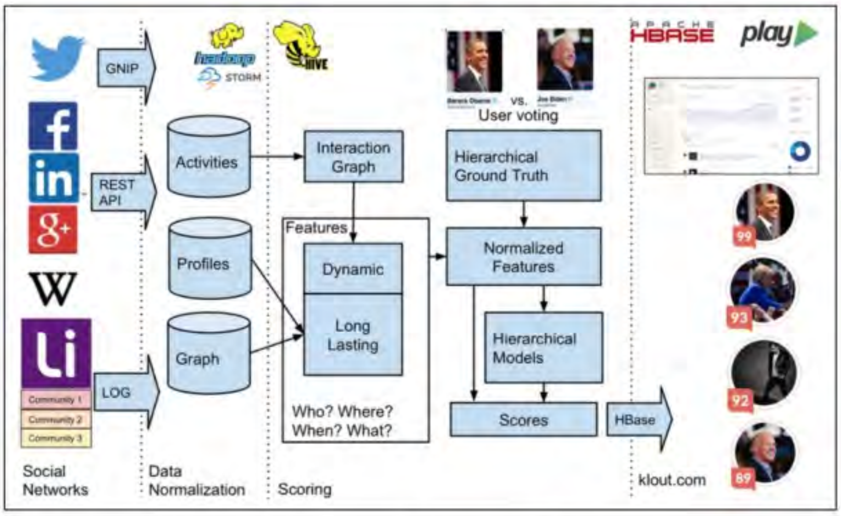
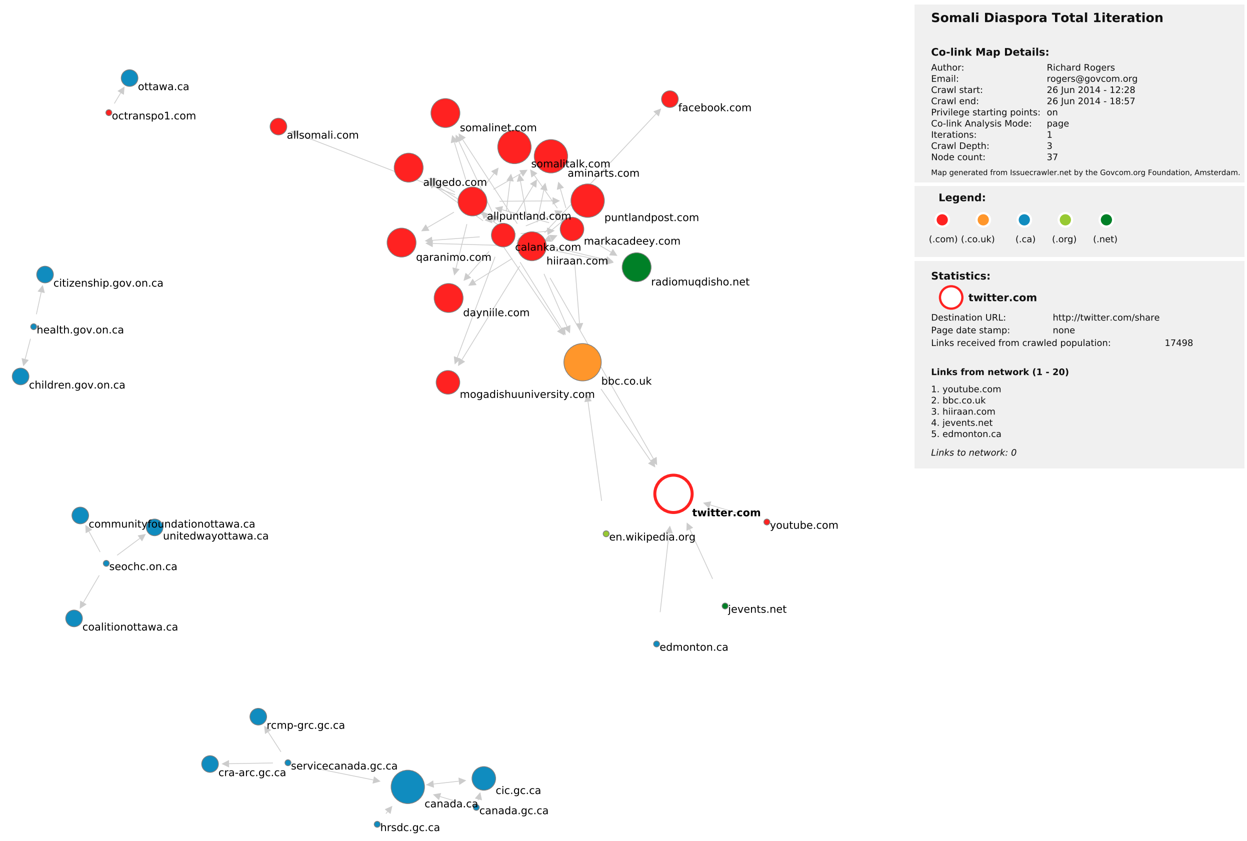


Figure Eight: Klout scoring mechanism as flow chart. Source: Rao, 2015.

When can so-called info-web methods based on the hyperlink still be applied to the study of the web and its platforms? By 'http or html approaches' to web 1.0 is meant software like the IssueCrawler and other hyperlink analysis tools, which generally speaking crawl a seed list of websites, locate hyperlinks either between them or between them and beyond them, and map the interlinkings, showing uni-directional, bi-directional as well as the absence of linking between websites (see figure nine). Problems arise. Through automated hyperlink analysis, the researcher may miss relationships between websites which are not captured by hyperlinks, such as sites mentioning each other in text without linking.

One may also miss links between websites because servers are down, or javascript or other code impenetrable to crawlers are employed on one or more websites in the network. (Elmer and Langlois thereby proposed to follow keywords across websites as well as platforms.)



Somali Diaspora Total 1iteration

Co-link Map Details:

Author: Richard Rogers
Email: rogers@govcom.org
Crawl start: 26 Jun 2014 - 12:28
Crawl end: 26 Jun 2014 - 18:57
Privilege starting points: on
Co-link Analysis Mode: page
Iterations: 1
Crawl Depth: 3
Node count: 37

Map generated from Issuecrawler.net by the Govcom.org Foundation, Amsterdam.

Legend:

● (.com) ● (.co.uk) ● (.ca) ● (.org) ● (.net)

Statistics:

○ **twitter.com**

Destination URL: http://twitter.com/share
Page date stamp: none
Links received from crawled population: 17498

Links from network (1 - 20)

1. youtube.com
2. bbc.co.uk
3. hiiraan.com
4. jevents.net
5. edmonton.ca

Links to network: 0

Figure Nine: IssueCrawler map showing Twitter.com as significant node, albeit without showing individual, significant Twitter users. Source: Issuecrawler.net, June 2014.

As the info-web has evolved into a social web, hyperlink analysis tends to continue to capture links between pages or hosts on the web, but not on social media platforms, where only the host is returned (Facebook.com), not individual user profiles (such as a Facebook account, page or group) or an individual Twitter user. (Similarly, Google's Web Search continually experiments with returning Twitter and Facebook content, however much web content remains privileged.) These drawbacks have occasioned researchers to move in two directions at once: develop crawlers and hyperlink analytical machines that pinpoint deep links between social media platforms and websites as well as within platforms (such as the Hyphe project[2]), and also to consider new means to study relationships between platforms as well as between platforms and the web that do not rely on hyperlinks only. Joining in part with the call by Elmer and Langlois, here the proposal would be to study content across the platforms (and the web): which content is co-linked, inter-liked and cross-hashtagged?

Co-linked content are URLs (often shortened on social media) that are linked by two or more users, platform pages or webpages. Inter-liked content is content liked by users and pages across platforms. Cross-hashtagged content is content referred to by hashtags across platforms.

[2] See the Hyphe project at the MediaLab, Sciences Po, Paris, http://hyphe.medialab.sciences-po.fr/.

Table Two. Elements of Cross-Platform Analysis

	Twitter	Facebook	Instagram
Query design	Hashtag(s), keyword(s), location(s), user(s)	Group(s), page(s)	Hashtag(s), location(s)
Data capture	In advance (for overtime data); on demand (for very recent data)	On demand (for overtime and recent data)	On demand (for overtime location data and recent hashtag data)
Platform user accounts (with primary actions)	user (follow)	user (friend, follow), group (join), page (like)	user (follow)
Content (media contents and digital objects)	tweet (text, photo, video, hashtag, @mention, URL, geotag)	post (text, video, photo, URL)	photo, video (text, hashtag, geotag, @mention)
Activities (resonance measures)	like (fav), retweet	like, comment, share	like, comment

Adapted from Rieder, 2015.

Research strategies for cross-platform analysis

How to perform cross-platform analysis? And which platforms may be productively compared. When discussing the kind of research done with social media, even with the shift to the study of social causes over the self, it is worthwhile to point out that one may emphasise medium research, social research or a combination of the two. For medium research, the question concerns how the platform affects the content, be it its presence or absence as well as its orderings. Additionally, specific cultures of use per platform, and (strategic) transmedia deployment, may inform the medium research, as discussed above. For social research, the question concerns the story the content tells, despite the platform effects. For a combination of medium and social research, the questions are combined; how does the platform affect the availability of content, and what stories do the content tell, given platform effects? Thus for cross-platform analysis, the following steps may be taken.

1) Choose a contemporary issue (revolution, disaster, election, social cause and so forth) for cross-platform analysis. One may choose to follow an active or unfolding issue (an issue in motion, so to speak), or one from recent history (an issue from the past, where overtime analysis is desirable). Here one should consider which platforms provide overtime data (Facebook), and which do not without great effort (Twitter).

2) Design a query strategy. For social issues and causes, consider querying for a program and an anti-program (see figure ten) (Rogers, 2016). For example, in the 2015 U.S. Supreme Court ruling for same-sex marriage the competing Twitter and Instagram hashtags reflected hashtag publics forming around a program and an anti-program, #lovewins and #jesuswins. If hashtags are preferred, for an election, consider querying a set of candidates or parties, e.g., #Trump and #Hillary (perhaps together with additional hashtags as well as keywords). For a disaster (or tragedy), consider querying its name(s), e.g., #MH17. URLs and/or domain names can be used as queries for a number of platforms.

3) Develop analytical strategy. For social issues and causes, consider which program or anti-program is finding favour (including amongst whom and where). Does it have a particular geography? For an election, consider creating portrayals of the candidates via the associated issues, or comparing their relative resonance with current election polls. For a revolution, consider its momentum and durability (including the subjects that continue to matter and those who do not endure). For a disaster, consider how it is (continually) remembered or forgotten, and to which extent it has been and still is addressed and by whom.

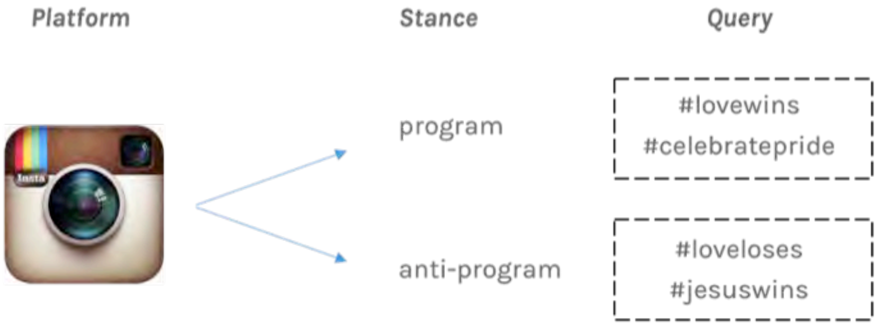


Figure Ten: Instagram query design strategy for the study of the images (and its geographies) associated with the U.S. Supreme Court ruling on same-sex marriage, 26 June 2015. Query design by participants of Digital Methods Summer School 2015, Does love win? The mechanics of memetics, https://wiki.digitalmethods.net/Dmi/SummerSchool2015DoesLoveWin.

4) Consider the configuration of use. It may be instructive for the analysis to look into how the platform is configured and set up by the initiator(s). Is it a group or a page, with or without moderation? Is it centrally organised or a collective effort? Are comments allowed? Does the user have a distinctive follower strategy?

5) Cross-platform analysis. Undertake the platform analysis, according to the query design strategy as well as the analytical strategy discussed above, across two or more platforms. For each platform consider engagement measures, such as the sum of likes, shares, comments (Facebook), likes and retweets (Twitter) and co-hashtags (Instagram). Which (media) content resonates on which platforms? Consider which content is shared across the platforms (co-linked, inter-liked and cross-hashtagged), and which is distinctive, thereby enabling both networked platform content analysis as well as medium-specific (or platform-specific) effects.

6) Discuss your findings with respect to medium research, social research or a combination of the two. Does a particular platform tend to host as well as order content in ways distinctive from other platforms? Are the accounts of the events distinctively different per platform or utterly familiar no matter the platform?

In practice certain platforms lend themselves to comparison more artfully than others, given both the availability of objects such as the hashtag or geotag as well as roughly similar cultures of use. Through the vehicle of the hashtag, Twitter and Instagram (as well as Tumblr) are often the subject of cross-platform analysis. One queries the APIs with such tools as TACT (for Twitter) as well as relatively simple Instagram and Tumblr hashtag explorers made available by the Digital Methods Initiative, creating collections of tweets and posts for further quantitative and qualitative analysis. Take for example certain significant events in the so-called migration crisis in Europe, one concerning the death of refugee children (Aylan Kurdi and his brother) and another the sexual assaults and rapes on New Year's Eve in Cologne (Geboers et al., 2016). For each case Twitter and Instagram are queried for hashtags (e.g., #aylan), whereupon tweet and post collections are made. For Twitter one 'recipe' to sort through the contents of the collections would include the following:

a) Hashtag Frequency counts ascertain the other hashtags that co-occur, and is useful to explore the issue space. For the Cologne rape cases, the hashtag #einearmlänge co-occurs greatly, which was a trending topic referring to the remarks by the Cologne mayor that (as a solution) women should remain an arm's length away from so-called strangers.
b) Mention Frequency lists the usernames of those who tweet and who are mentioned so one note which users may dominate a space.
c) Retweet Frequency provides a ranked list of retweeted tweets, showing popular or significant content.
d) URL Frequency is a ranked URL lists showing popular or significant media (such as images and video). The most referenced media, especially images, become a focal point for a cross-platform analysis with Twitter.

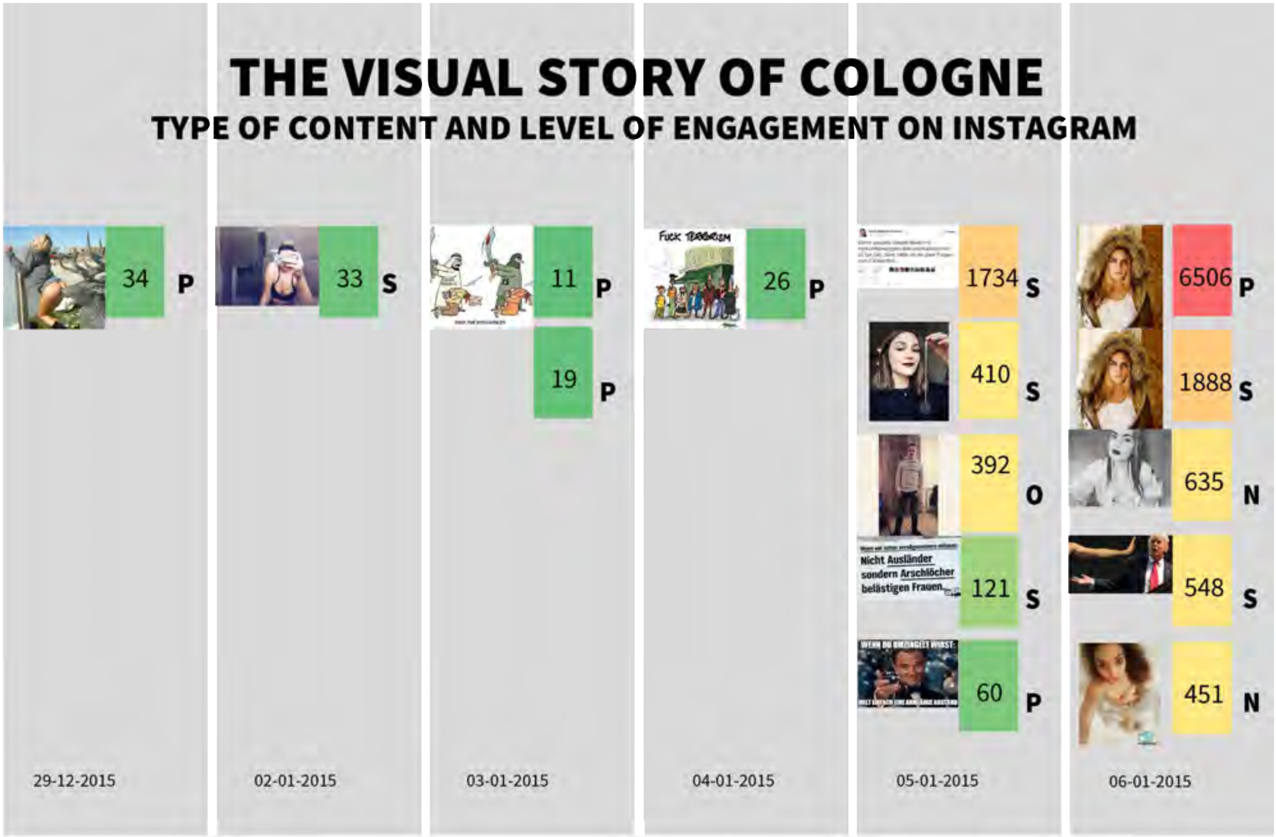
For Instagram, hashtag frequency is undertaken together with image and video frequency. (One is also able to query Instagram for geo-coordinates, which is not undertaken here.) Ultimately, the means of comparison are hashtag as well as image and video use, where the former suffers somewhat from hashtag stuffing in Instagram.

The question of platform effects is treated in the qualitative analysis, where in both cases the incidence of news photos was much greater in Twitter than in Instagram, where there were more derivatives, meaning annotated, photoshopped, cartoon-like or other DIY materials with (implied or explicit) user commentary. Twitter thereby becomes a professional medium (with effects) and Instagram more a user-generated content medium. The Aylan case, however, appears to reduce this medium-specificity, because there is a relatively greater amount of images which have been edited so as to come to grips with the tragedy of the drowned toddler (see figures 11A-D).

Conclusions: Digital Methods for Cross-platform Analysis

In the call for methodological attention to the platformisation of the web, Langlois and Elmer discuss how analyses based on the hyperlink do not embrace the analytical opportunities afforded by social media. Hyperlink analysis, and its tools such as the Issuecrawler, rely on an info-web (aka web 1.0), where webmasters make recommendations by linking to another website (or non-recommendations through not making links, thereby showing lack of interest or affiliation). Focusing on links only misses the novel objects of web 2.0, social networking sites, platforms and social media (as the social web has been called), such as the like, share and tweet. Whilst Langlois and Elmer called for the analysis of the keyword both over the hyperlink but also perhaps over other social media objects, around the same time as their publication the API had arrived (Facebook's version 1.0 in 2010, Twitter's in 2006), and gradually became the preferred point of access to data over scraping which the platforms actively sought to thwart. The API is of course controlled by the service in question, be it Twitter, Facebook or others, and steers research in ways more readily palpable perhaps than scraping, for the data available on the interface (that could be scraped) and through the developer's entry point may differ considerably. The ethics turn in web research, bound up with the rise of the social web and its publicly available, personal data, in turn has shaped the accessibility of certain data on the APIs such that Facebook no longer allows one to collect friends' 'tastes and ties', or likes, profile interests as well as friends. Such unavailability comes on the heels of a critique of a study of the same name that collected (or scraped, albeit with permission) Facebook profiles and friends data from Harvard students and enriched it with their student housing information, without their knowledge. Concomitant with the decline in the study of the self in social media (given the increasing dearth of available data) has been the rise in attention to events, disasters, elections, revolutions and social causes. Not only is it in evidence in Facebook research on (Arab Spring) pages (and to an extent groups), but also in Twitter (revolutions), where Jack Dorsey, its co-founder, signalled the shift in the interviews in the Los Angeles Times in 2009, mentioning that Twitter did well events such as disasters, elections as well as conferences. Instagram, according to its founder Kevin Systrom, would like to follow the same trajectory becoming a platform of substance and thereby for the study of events. The API, however, appears to have shaped social media studies beyond its selective availability of data. Rather the APIs serve as silos for what I call 'single-platform studies', which are reflected in the available tools discussed. Netvizz is for Facebook studies, TCAT for Twitter studies, the Instagram hashtag explorer for Instagram and so forth. Unlike the web 1.0 tools such as Issuecrawler, which find links between websites and between websites and platforms, the social web has not seen tools developed for cross-platform analysis. Where to begin?

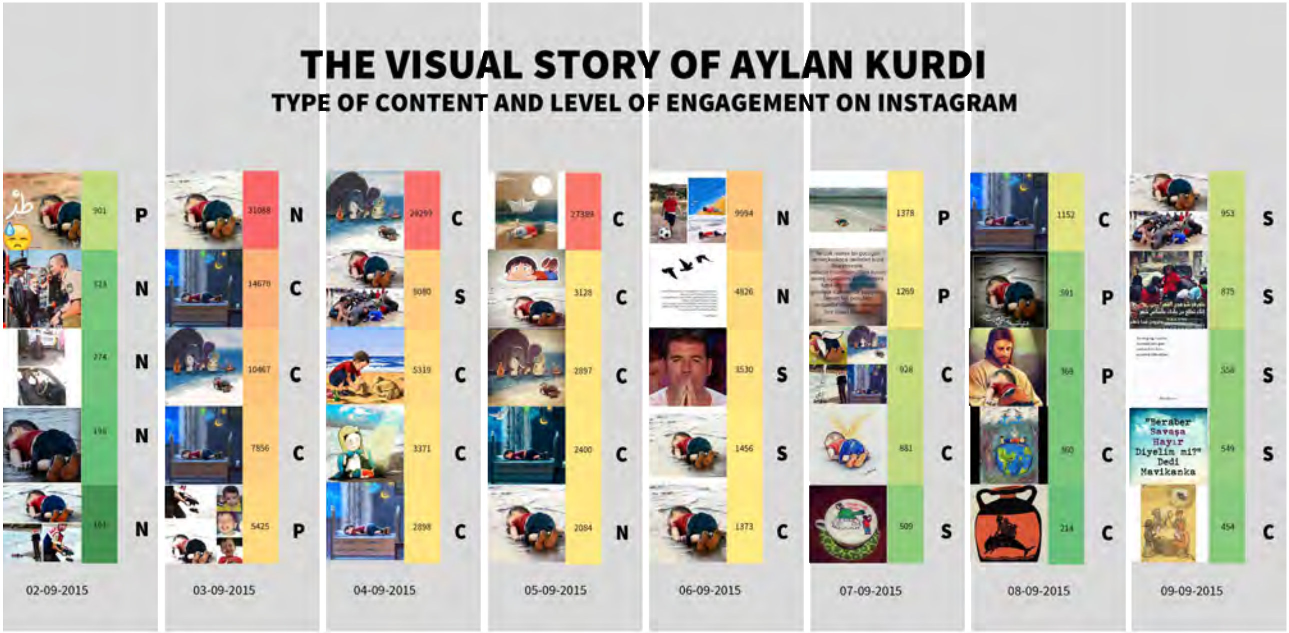
The purpose here is to develop techniques for multiple platform analysis that bear medium-sensitivity. Stock is taken of the objects that platforms share, whereupon cultures of use are taken into consideration. In other words, Twitter, Facebook and Instagram share the hashtag, however much on the one no more than two are recommended, on another it is rarely used and on the third it is used in overabundance. The cross-platform approaches that are ultimately described rely on hashtags for making collections of tweets (in Twitter) and posts (in Instagram), whereupon the media format (images, but also videos) common to the two are compared in the study of events. During the European refugee crisis of 2015-2016 the death of the toddler, Aylan Kurdi, and the sexual assaults of women in Cologne, stand out as major (social media) events for analysis with a quanti-quali approach and a networked content analysis, which are forms of analysis with affinities with computational hermeneutics.



THE VISUAL STORY OF COLOGNE

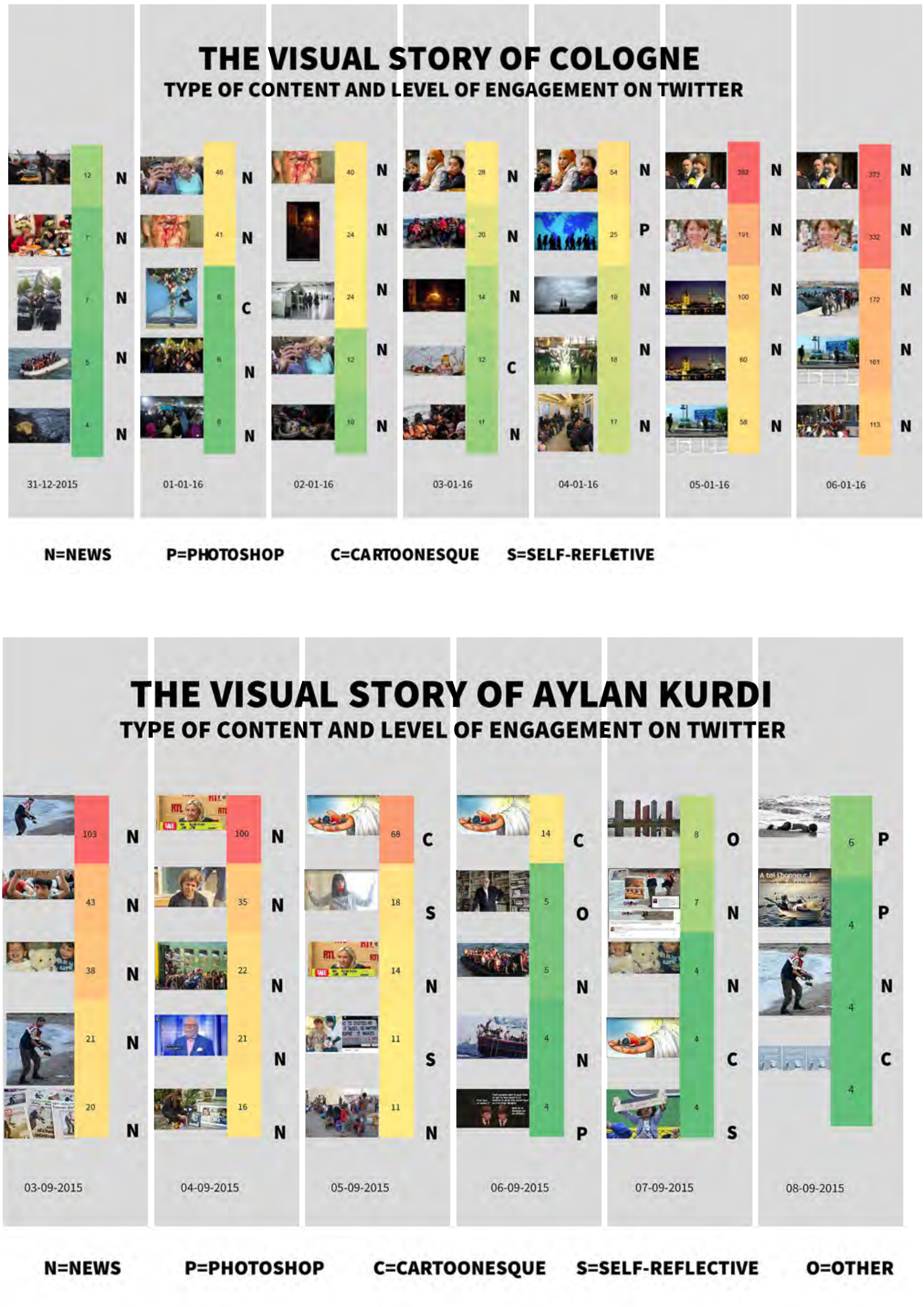
TYPE OF CONTENT AND LEVEL OF ENGAGEMENT ON INSTAGRAM

N=NEWS P=PHOTOSHOP C=CARTOONESQUE S=SELF-REFLECTIVE



THE VISUAL STORY OF AYLAN KURDI

TYPE OF CONTENT AND LEVEL OF ENGAGEMENT ON INSTAGRAM



Figures Eleven A-D: Most frequently occurring images in Twitter and Instagram for the Aylan and Cologne rape cases in the European refugee crisis, 2015-2016, categorised by image type. Source: Geboers et al., 2016.

Suggested resources

For video tool tutorials, see the DMI ‘tools walkthrough’ playlist on YouTube, https://www.youtube.com/playlist?list=PLKzQwIKtJvv9IwyYxh4708Nqo6YC6-YH4

1) Instagram

Instagram software tools
- Instagram hashtag explorer
- Instagram network
- Instagram scraper
http://tools.digitalmethods.net

Video tutorial for Instagram hashtag explorer, “Analyze Instagram Activity Around a Hashtag or Location”
https://www.youtube.com/watch?v=o07aUKdRv0g

2) Twitter

DMI-TCAT (Twitter Capture and Analysis Tool)
http://tools.digitalmethods.net

List of all datasets currently captured by
https://tools.digitalmethods.net/beta/tcat/

Video tutorial for TCAT, “Overview of Analytical Modules”
https://www.youtube.com/watch?v=ex97eoorUeo

3) Facebook

Netvizz (Facebook Data Extraction Tool)
http://tools.digitalmethods.net

Netvizz video tutorials:
“Introduction to Netvizz 1.2+”
https://www.youtube.com/watch?v=3vkKPcN7V7Q

“Downloading data and producing a macro view”
https://www.youtube.com/watch?v=dfoYAPistYg

4) Gephi-related

Gephi (The Open Graph Viz Software)
https://gephi.org

“Gephi Tutorial for working with Twitter mention networks”
https://www.youtube.com/watch?v=snPR8CwPId0

“Combine and Analyze Co-Hashtag Networks (Instagram, Twitter, etc.) with Gephi” https://www.youtube.com/watch?v=ngqWjgZudeE

References

John Agar (2001). “Review of James Gillies and Robert Cailliau, How the Web was Born. Oxford: Oxford University Press, 2000,” The British Journal for the History of Science 34(3): 370-373.

Michael Barbaro and Tom Zeller Jr. (2006). “A Face is Exposed for AOL Searcher no. 4417749,” *The New York Times.* pA1.

David Beer (2008). “Social Network(ing) Sites...Revisiting the Story so far: A Response to danah boyd & Nicole Ellison,” *Journal of Computer-Mediated Communication*. 13(2): 516–529.

W. Lance Bennett & Alexandra Segerberg (2012). “The Logic of Connective Action: Digital media and the personalization of contentious politics,” *Information, Communication & Society*. 15(5): 739-768.

Andrew 'Boz' Bosworth (2009). “What's the history of the Awesome Button (that eventually became the Like button) on Facebook?,” *Quora*, https://www.quora.com/Whats-the-history-of-the-Awesome-Button-that-eventually-became-the-Like-button-on-Facebook.

danah boyd and Nicole Ellison (2007). “Social Network Sites: Definition, History and Scholarship,” *Journal of Computer-Mediated Communication*, 13(1), article 1.

Petter Bae Brandtzaeg and Ida Maria Haugstveit (2014). “Facebook likes: A study of liking practices for humanitarian causes,” *International Journal of Web Based Communities*. 10(3): 258-279.

Axel Bruns and Jean Burgess (2015). “Twitter Hashtags from Ad Hoc to Calculated Publics,” in Nathan Rambukkana (ed.) *Hashtag Publics: The Power and Politics of Discursive Networks*. New York: Peter Lang, 13-28.

Axel Bruns, Jean Burgess, and Tim Highfield (2014). “A ‘Big Data’ Approach to Mapping the Australian Twittersphere.” In Paul Longley Arthur and Katherine Bode, eds., Advancing Digital Humanities: Research, Methods, Theories. Houndmills: Palgrave Macmillan, 113-129.

danah boyd and Eszter Hargittai (2010). “Facebook privacy settings: Who cares?” *First Monday*. 15(8).

Taina Bucher (2013). Objects of Intense Feeling: The Case of the Twitter API,” *Computational Culture: A Journal of Software Studies*, 4.

Henrik Serup Christensen (2011). “Political activities on the Internet: Slacktivism or political participation by other means?” *First Monday*. 16(2).

Sasha Constanza-Chock (2014). *Out of the Shadows, Into the Streets! Transmedia Organising and the Immigrant Rights Movement*. Cambridge, MA: MIT Press.

Annet Dekker and Annette Wolfsberger (2009). *Walled Garden*. Amsterdam: Virtual Platform.

Greg Elmer and Ganaele Langlois (2013). "Networked Campaigns: Traffic Tags and Cross Platform Analysis on the Web," *Information Polity*. 18(1): 43–56.

Lernert Engelberts and Sander Plug (2009). "I Love Alaska: The Heartbreaking Search History of AOL User #711391," Minimovies Documentary. Amsterdam: Submarine Channel.

Facebook (2016). "Facebook Platform Changelog," Facebook for Developers, webpage, https://developers.facebook.com/docs/apps/changelog.

Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong Devin Gaffney (2010). "#iranElection: Quantifying online activism," *Proceedings of WebSci10*. New York: ACM.

Marloes Geboers, Jan-Jaap Heine, Nienke Hidding, Julia Wissel, Marlie van Zoggel and Danny Simons (2016). "Engagement with Tragedy in Social Media," Digital Methods Winter School '16, Amsterdam, https://wiki.digitalmethods.net/Dmi/WinterSchool2016EngagementWithTragedySocialMedia.

Carolin Gerlitz and Anne Helmond (2013). "The Like Economy: Social Buttons and the Data-intensive Web," *New Media & Society*. 15(8): 1348-1365.

Tarleton Gillespie (2010). "The politics of 'platforms'," *New Media & Society*. 12(3): 347-364.

Vindu Goel (2015). "Instagram to Offer Millions of Current Events Photos, *New York Times*, 23 June.

Anne Helmond (2015). *The Web as Platform: Data Flows in Social Media*. PhD dissertation. University of Amsterdam.

Henry Jenkins (2006). *Convergence Culture*. New York: New York University Press.

Karrie Karahalios, Kevin Hamilton and Christian Sandvig (2015). "'I always assumed that I wasn't really that close to [her]': Reasoning about Invisible Algorithms in News Feeds." *CHI 2015, Crossings*, Seoul, South Korea. New York: ACM.

Adam D.I. Kramer, Jamie E. Guilloryb and Jeffrey T. Hancock (2014). "Experimental evidence of massive-scale emotional contagion through social networks," *PNAS*. 111(24): 8788–8790.

Akarshan Kumar, "Hearts on Twitter (and Vine)," *Twitter blog*, 3 November 2015, https://blog.twitter.com/2015/hearts-on-twitter.

Kevin Lewis, Jason Kaufman, Marco Gonzalez, Andreas Wimmer and Nicholas Christakis (2008). "Tastes, Ties, and Time: A New Social Network Dataset using Facebook.com," *Social Networks*. 30(4): 330–342.

Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, danah boyd (2011). "The Arab Spring | The Revolutions Were Tweeted: Information Flows during the 2011

Tunisian and Egyptian Revolutions," *International Journal of Communication*, 5.

Noortje Marres and Esther Weltevrede (2013). "Scraping the Social? Issues in live Social Research," *Journal of Cultural Economy*. 6(3): 313-335.

Michelle N. Meyer (2015). "Two Cheers for Corporate Experimentation: The A/B Illusion and the Virtues of Data-Driven Innovation," 13 Colo. Tech. L.J. 273, http://ssrn.com/abstract=2605132.

John W Mohr, Robin Wagner-Pacifici, Ronald L Breiger (2015). "Toward a Computational Hermeneutics," *Big Data & Society*. 2(2).

Sabine Niederer (2016). *Networked Content Analysis: The Case of Climate Change*. PhD dissertation. University of Amsterdam.

Helen Nissenbaum (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, CA: Stanford University Press.

Tim O'Reilly (2005). "What is Web 2.0: Design Patterns and Business Models for the NextGeneration of Software," O'Reilly, Sebastopol, CA: O'Reilly Media, http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html.

Zizi Papacharissi (2015). *Affective Publics*. New York: Oxford University Press.

Cornelius Puschmann and Engin Bozdag (2014). "Staking out the unclear ethical terrain of online social experiments," *Internet Policy Review*. 3(4).

Adithya Rao, Nemanja Spasojevic, Zhisheng Li and Trevor DSouza (2015). "Klout Score: Measuring Influence Across Multiple Social Networks," *2015 IEEE International Big Data Conference - Workshop on Mining Big Data in Social Networks*. New York: ACM.

Bernhard Rieder (2012). What is PageRank? A Historical and Conceptual Investigation of a Recursive Status Index," *Computational Culture*, 2.

Bernhard Rieder (2013). "Studying Facebook via data extraction: the Netvizz application," *Proceedings of the 5th Annual ACM Web Science Conference*. New York: ACM Press, 346-355.

Bernhard Rieder (2015). "Social Media Data Analysis," lecture delivered at the University of Amsterdam, December.

Bernhard Rieder, Rasha Abdulla, Thomas Poell, Robbert Woltering and Liesbeth Zack (2015). "Data critique and analytical opportunities for very large Facebook Pages: Lessons learned from exploring "We are all Khaled Said," *Big Data & Society*. 2(2): 1–22.

Piers Robinson (2002). *The CNN Effect: The Myth of News, Foreign Policy and Intervention*. London: Routledge.

Richard Rogers (2009). "Post-demographic Machines," in Annet Dekker and Annette Wolfsberger (eds.), *Walled Garden*. Amsterdam: Virtual Platform, 29-39.

Richard Rogers (2013). "Debanalizing Twitter: The Transformation of an Object of Study," *Proceedings of WebSci13*. New York: ACM.

Richard Rogers (in press). "Foundations of Digital Methods: Query Design," in Mirko Schaefer and Karin van Es (eds.), *The Datafied Society: Studying Culture Through Data*.

Amsterdam: Amsterdam University Press.

Richard Rogers, Esther Weltevrede, Erik Borra, Marieke van Dijk and the Digital Methods Initiative (2009). “For the ppl of Iran: #iranelection RT,” in Gennaro Ascione, Cinta Massip and Josep Perello (eds.), *Cultures of Change: Social Atoms and Electronic Lives*. Barcelona: Actar and Arts Santa Monica, 112-115.

Richard Rogers, Esther Weltevrede, Sabine Niederer and Erik Borra (2013), “National Web Studies: The Case of Iran Online,” in John Hartley, Axel Bruns, and Jean Burgess (eds.), *A Companion to New Media Dynamics*, Oxford: Blackwell, 142-166.

David Sarno (2009). “Jack Dorsey on the Twitter ecosystem, journalism and how to reduce reply spam: Part II,” *Los Angeles Times*, 19 February.

Theresa Senft and Nancy Baym (2015). “What Does the Selfie Say? Investigating a Global Phenomenon,” International Journal of Communication 9: 1588–1606.

Fred Stutzman, Ralph Gross and Alessandro Acquisti (2013). "Silent Listeners: The Evolution of Privacy and Disclosure on Facebook," Journal of Privacy and Confidentiality. 4(2), article 2.

Twitter (2016). “Using hashtags on Twitter,” Help Center, San Francisco: Twitter, Inc., https://support.twitter.com/articles/49309

Elizabeth Van Couvering (2007). “Is Relevance Relevant? Market, Science, and War: Discourses of Search Engine Quality,” *Journal of Computer-Mediated Communication*. 12(3), article 6, http://jcmc.indiana.edu/vol12/issue3/vancouvering.html.

Ibo Van de Poel (2009). “The Introduction of Nanotechnology as a Societal Experiment,” in Simone Arnaldi, Andrea Lorenzet and Federica Russo (eds.), *Technoscience in Progress. Managing the Uncertainty of Nanotechnology.* Amsterdam: IOS Press, 129–142.

Tommaso Venturini, Dominique Cardon and Jean-Philippe Cointet (2014), “Présentation - Méthodes digitales: Approches quali/quanti des données numériques,” *Réseaux.* 188(6), 9-21.

Esther Weltevrede (2016). *Repurposing Digital Methods: The Research Affordances of Engines and Platforms*. PhD dissertation. University of Amsterdam.

Michael Zimmer (2008). “More on the ‘Anonymity’ of the Facebook Dataset — It’s Harvard College,” Michaelzimmer.org blog, http://michaelzimmer.org/2008/10/03/moreon-the-anonymity-of-the-facebook-dataset-its-harvard-college/.

Michael Zimmer (2010). “But the Data is already Public: On the Ethics of Research in Facebook,” *Ethics and Information Technology*. 12(4): 313-325.

To appear in Shaefer, M. et al. (eds.), *The Datafied Society*. Amsterdam: Amsterdam University Press, 2016

Foundations of Digital Methods: Query Design

Richard Rogers

Digital methods and online groundedness

Broadly speaking digital methods may be considered the deployment of online tools and data for the purposes of social and medium research. More specifically, they derive from online methods, or methods of the medium, which are reimagined and repurposed for research. The methods to be repurposed are often built into dominant devices for recommending sources or drawing attention to oneself or one's posts. For an example of how to reimagine the inputs and outputs of one such dominant device, consider the difference between studying search engine results to understand in some manner Google's algorithms, or recent algorithmic updates, or treating them, as in the Google Flu Trends project, as indications of societal concerns. Here, there is a shift from studying the medium to using device data to study the societal. That is, akin to the digital methods outlook generally, Google Flu Trends and other anticipatory instruments use online social signals to measure trends not so much in the online realm but rather 'in the wild'. [1]

Once the findings are made the question becomes how to ground them, that is, with conventional offline methods and techniques, such as the Center for Disease Control's means of studying flu incidence through hospital and doctor reports, as in the Flu Trends project, or through additional, online methods and sources. In digital methods research, online groundedness, as I have called it, asks whether and when it is appropriate to shift the site of 'ground-truthing', to use a geographer's expression. As a case in point, when verifying knowledge claims, Wikipedians check prior art through Google searches, thereby grounding claims via the search engine in online sources.

Digital methods thereby rethink conditions of proof, first by considering the online as a site of grounding, but also in a second sense. One makes social research findings online, and, rather than leaving the medium to harden them, one subsequently inquires into the extent to which the medium is affecting the findings. Medium research thus serves a purpose that is distinct from the study of online culture alone. As I come to shortly, when reading and interpreting social signals online, the question concerns whether the medium, or media dynamics, are over determining the outcomes.

Making use of online data: From the semantic to the social

As noted digital methods make use of online methods, by which I refer to an array of techniques from the computational and information sciences -- crawling, scraping, indexing, ranking and so forth -- that have been applied to and redeveloped for the web. They refer to algorithms that determine relevance and authority and thereby recommend information sources as in Google's famed PageRank, but also boost all manner of items, from songs and 'friends' to potential 'followers'.

Many of the algorithms are referred to as 'social', meaning that they make use of user choices and activity (purposive clicks such as liking), and may be contrasted with the 'semantic', meaning that which is categorised and matched (as in Google's Knowledge Graph). Digital methods seek to take particular advantage of socially derived rankings, that is, users making their preferences known for particular sources, often unobtrusively. Secondarily, the semantic (sources that have been pre-matched or taxonomied) are also of value, for example when Wikipedia furnishes a curated seed list of sources ("climate change sceptics," as a case in point), which have been derived manually by information experts or the proverbial crowd guided by the protocols of the online encyclopaedic community.

The distinction between social and semantic is mentioned so as to emphasize web-epistemological 'crowdfindings' (as implied by the 'social'), as distinct from 'results' from information retrieval.[2] Thus with digital methods, as I relate below, one seeks to query in order to make findings from socialised web data (so to speak) rather than query in order to find pre-sorted information or sources, however well annotated or enriched with meta-data.

Why query Google (still) for research purposes?

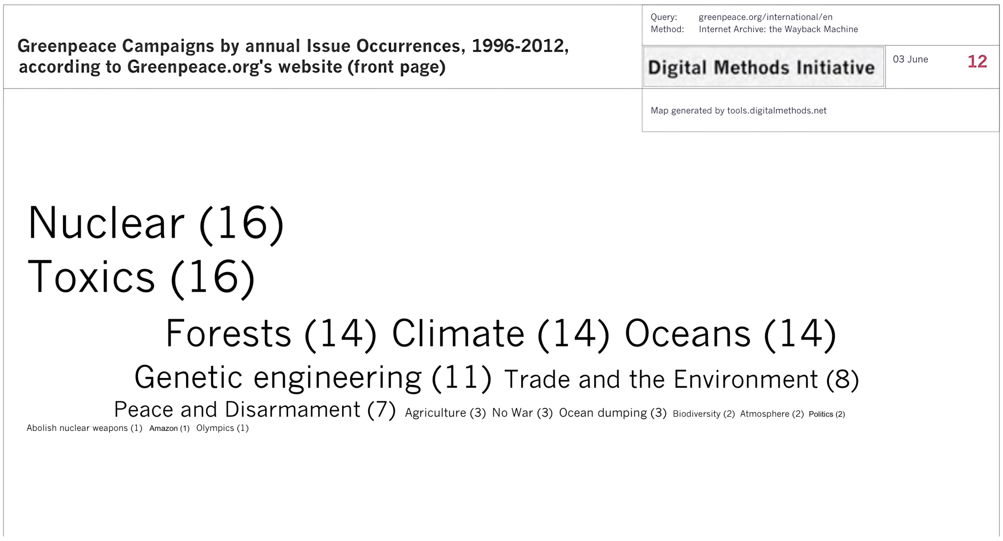
Over the course of the past decade or more Google arguably has transformed itself from an epistemological machine outputting reputational source hierarchies to a consumer information appliance providing user-tailored results. Here I would like to take up the question of how and to which ends one might still employ Google as an epistemological machine.

There are largely two research purposes for querying Google: medium and social research. With medium research, one studies (often critically) how and for whom Google works. To which degree does the engine serve a handful of dominant websites such as Google properties themselves in a 'preferred placement' critique, or websites receiving the most attention through links and clicks? One would seek to lay bare the persistence of so-called "googlearchies" that boost certain websites and bury others in the results, as Matthew Hindman's classic critique of Google's outputs would imply. Here the work being done is an engine results critique, where the question revolves around the extent to which the change in 2009 in Google's algorithmic philosophy, captured in the opening chapter of Eli Pariser's *Filter Bubble,* from universal to personalised outputs, dislodges or upholds the pole positions of dominant sites on the web. Indeed, another critical inroad in engine results critique is the so-called filter bubble itself, where one would examine the effects of personalisation, investigating Pariser's claim that Google furnishes increasingly personalised and localised results. In this enquiry, one may reinvigorate Nicholas Negroponte's 'Daily Me' argument and Cass Sunstein's response concerning the undesirable effects of homophily, polarization, and the end of the shared public exposure to media which leaves societies without common frames of reference. In this line of reasoning, personalisation leads to social atomisation and severe niching, otherwise known as 'markets of one', as described by Joseph Turow in *Niche Envy*. It also would imply the demise of the mass media audience.

In the second research strategy, there is a mode switch in how one views the work of the search engine (and for whom it could work). Google's queries, together with its outputted site rankings, are considered as indicators of societal trends. That is, instead of beginning from the democratizing and socializing potential of the web and subsequently critiquing Google for its reintroduction of hierarchies, one focuses on how examining engine queries and results allows for the study of social sorting. How to study the hierarchies Google offers? Which terms have been queried most significantly (at which time and from which location)? Do places have preferred searches? May we geo-locate temporal pockets of anxiety? The capacity to indicate general and localisable trends makes Google results of interest to the social researcher. [3]

Apart from trends one may also study dominant voice, commitment and concern. One may ask in the first instance, when and for which keywords do certain actors appear high on the list and others marginal? Which actors are given the opportunity to dominate and drive the meaning of terms and their discussion and debate? Here the engine is considered as serving social epistemologies for any keyword (or social issue) through what is collectively queried and returned.

The engine also can be employed to study of commitment in terms of the continued use of keywords by individual actors, be they governments, non-governmental organizations, radical group formations or persons. Here the researcher takes advantage not of the hierarchies inputted and outputted (socio-epistemological sorting) but of the massive and recent indexing of individual websites. For example, the non-governmental organization Greenpeace once had the dual agenda of environmentalism and disarmament (hence the fusion of 'green' and 'peace'). Querying Greenpeace websites lately for issue keywords would show that their commitment to campaigning for peace has significantly waned in comparison to that for environmental causes, for green words resonate far more than disarmament ones. Here one counts incidences of keywords on web pages for the study of issue commitment (see Figures 1 and 2).


Greenpeace Campaigns by annual Issue Occurrences, 1996-2012, according to Greenpeace.org's website (front page)

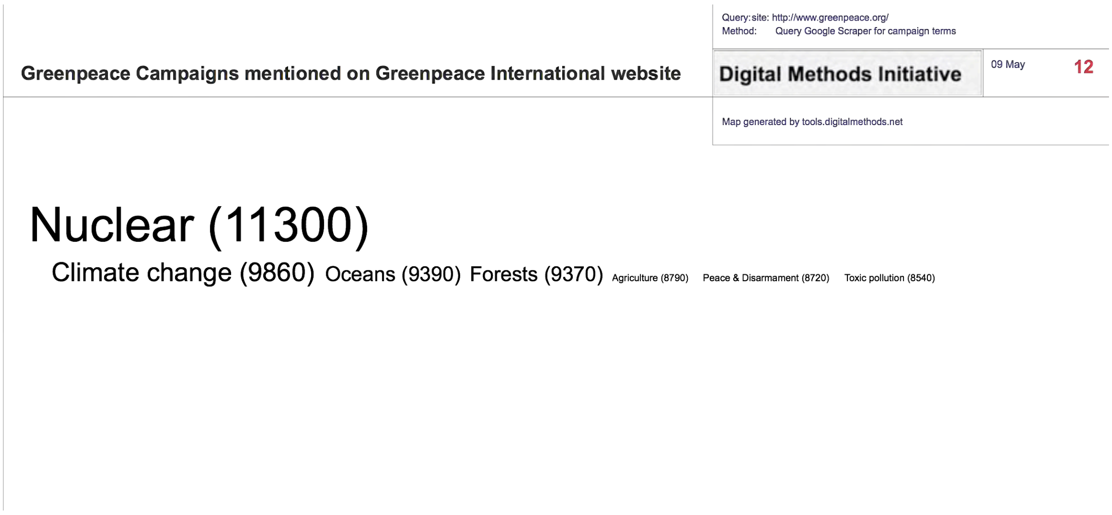
Query: greenpeace.org/international/en
Method: Internet Archive: the Wayback Machine

Digital Methods Initiative 03 June 12

Map generated by tools.digitalmethods.net

Nuclear (16)
Toxics (16)
Forests (14) Climate (14) Oceans (14)
Genetic engineering (11) Trade and the Environment (8)
Peace and Disarmament (7) Agriculture (3) No War (3) Ocean dumping (3) Biodiversity (2) Atmosphere (2) Politics (2)
Abolish nuclear weapons (1) Amazon (1) Olympics (1)

FIGURE 1. GREENPEACE CAMPAIGNS, 1996-2012, RANKED AND ARRAYED AS WORD CLOUD ACCORDING TO FREQUENCY OF APPEARANCES ON GREENPEACE.ORG FRONT PAGE. SOURCE: DATA FROM THE INTERNET ARCHIVE, ARCHIVE.ORG. ANALYSIS BY ANNE LAURINE STADERMANN.


Greenpeace Campaigns mentioned on Greenpeace International website

Query:site: http://www.greenpeace.org/
Method: Query Google Scraper for campaign terms

Digital Methods Initiative 09 May 12

Map generated by tools.digitalmethods.net

Nuclear (11300)
Climate change (9860) Oceans (9390) Forests (9370) Agriculture (8790) Peace & Disarmament (8720) Toxic pollution (8540)

FIGURE 2. GREENPEACE CAMPAIGNS MENTIONED ON GREENPEACE.ORG AS RANKED WORD CLOUD, 2012. SOURCE: DATA FROM GREENPEACE.ORG GATHERED BY THE LIPPMANNIAN DEVICE, DIGITAL METHODS INITIATIVE. ANALYSIS BY ANNE LAURINE STADERMANN.

One also may query sets of actors for keywords in order to have an indication of the levels of concern for an issue. For example, querying a representative environmental group and a species group (respectively) for Fukushima would show that the environmental group is highly active in the issue space whilst the species NGO is largely absent, showing a lack of concern for the matter (see Figure 3).

In all, for the social researcher, Google is of interest for its capacity to rank actors (websites) per social issue (keyword), thereby providing source hierarchies, and allowing for the study of dominant voice. It is also pertinent for its ability to count the incidence of issue words per actor or sets of actors, thereby allowing for the study of commitment through continued use of keywords.

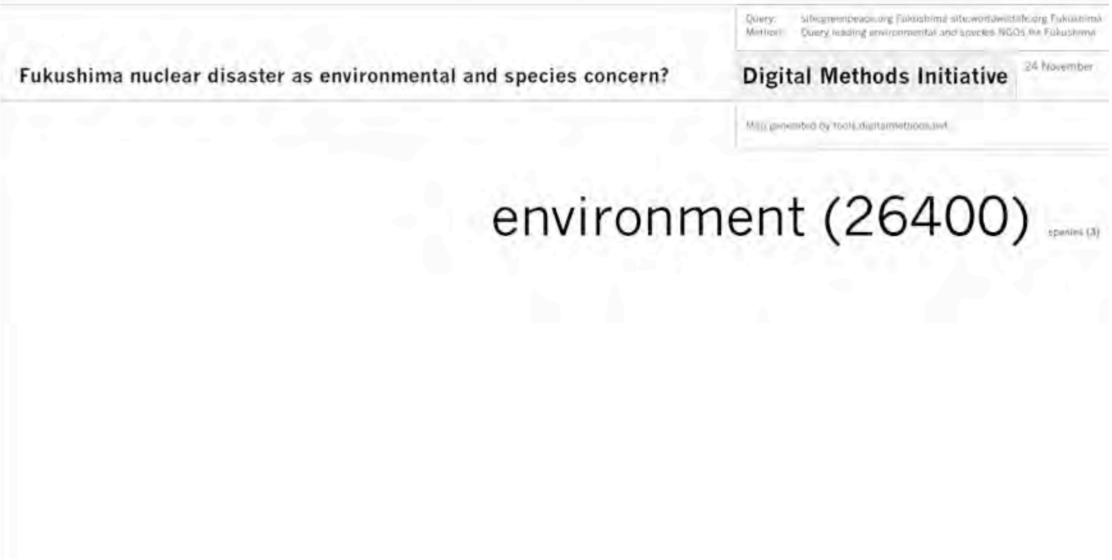


FIGURE 3. GREENPEACE WITH NUMEROUS MENTIONS OF FUKUSHIMA AND WORLD WILDLIFE WITH FEW, NOVEMBER 2016. SOURCE: DATA AND VISUALISATION BY THE LIPPMANNIAN DEVICE, DIGITAL METHODS INITIATIVE.

Clean Google results to remove 'artefacts'?

One might distinguish between the two research types above by viewing one as primarily doing media studies and the other social research. Yet in practice, the two are entangled with one another. As mentioned in the introduction here the entanglement assumes a particular form. Medium research is in service of social research in the sense of concentrating on the extent to which the findings made have been over determined by media effects.

It is important to stress from the outset that it not assumed that engine effects can be removed *in toto*, thus enabling a researcher to study 'organic' results, the industry term for editorial content untouched by advertising or preferred placement. Rather there should be awareness of a variety of types of routinely befouling artefacts ('media effects') that nevertheless are returned by the engine. Google properties (e.g., YouTube videos), Google user aids (e.g., 'equivalent results' for queried terms), and SEO'd products (whether through white or black hat techniques) are all considered media effects, and in principle could be removed, or footnoted. There are software settings (e.g., remove Google properties from results), query design (use quotation marks for exact matches) and also strategies for detecting at least obviously SEO'd results.

The more problematic issue arises with any desired detection of the effects of personalisation. The point here is that users now co-author engine results. The search engine thereby produces artefacts that are of the user's making. The search engine, once critiqued for its social sorting and Matthew effect in the results, leans towards inculpability, since users have set preferences (and had preferences set for them) and some results are affected. There is the question of detecting how many and which results are personalised in one form or another, according to one's location (country as well as locality), language, personal search history as well as adult and violent content filter.

Certain queries would likely have no organic results in the top ten, thus making any content cleaning exercise into an artificial act of removal, given that most users a) click the top results, b) have the results set to the default of ten, and c) do not venture beyond one page of results. There are also special cases to consider for removal, such as Wikipedia, which is delivered in the top results for nearly all substantive queries, making it appear to be at once an authoritative source (for its persistent presence) and an engine artefact (for its uncannily persistent presence). Wikipedia's supra-presence, so to speak, provides a conundrum for the researcher who may wish to clean content of Google artefacts and media effects, and is perhaps the best case for retaining them at least in the first instance.

One way forward would be to remove the user, so speak, and strive to have the engine work as unaffected as possible. Removing the user is a means of re-conjuring the pre-2009 distinction between universal results (served to all) and personalised results (served to an individual user). A research browser would be set up, where one is logged out of Google, and no cookies are set. The ncr (no country redirect) version of Google is used, or one would query from a non-location, or obfuscated one.

Studying media effects or the societal 'in the wild'?

The question of whether Google merely outputs Google artefacts and medium effects or reveals societal trends has been raised in connection with the flagship big data project, Google Flu Trends (Lazer et al., 2014). As mentioned at the outset, the project, run by Google's non-profit Google.org, monitors user queries for flu and flu-related symptoms, geolocates their incidence and outputs the timing and locations of heightened flu activity; it is a tool for tracking where the virus is most prevalent. Yet does the increased incidence of queries for flu and flu-related symptoms indicate a rise in the number of influenza cases 'in the wild', or does it mean that TV and other news of the coming flu season prompt heightened query activity? TV viewers may be using a 'second screen' and fact checking or enhancing their knowledge through search engine queries. Given that Flu Trends was over reporting for a period of time, compared to its baseline at the National Center for Disease Control (and its equivalents internationally), the project seemed to be overly imbued with media effects.

Thus one may seek research strategies to study medium effects, formulating queries that in a sense put on display or amplify the effects. For which types of queries do more Google properties appear? How can Google be made to output user aids that are telling? How to detect egregiously SEO'd results?

When using Google as a social research machine, the task at hand, however, is to reduce Google effects, albeit without the pretension of completely removing them. This is the main preparatory work, conceptually as well as practically, prior to query design.

When words are keywords: A query design strategy

The question of what constitutes a keyword is the starting point for query design for that is what makes querying and query design practically a part of a research strategy. When formulating a query, one often begins with keywords so as to ascertain who is using them, in which contexts and with which spread or distribution over time. In the following a particular keyword query strategy or design is put forward, whereby one queries competing keywords, asking whether a particular term is winning favour and amongst whom.

The keyword has its origins in the notion of a 'hint' or 'clue.' *New Oxford American Dictionary* (built into Apple OS's dictionary) calls it "a word which acts as the key to a cipher or code." In this rendering keywords do not so much have hidden but rather purposive meaning so as to enable an unlocking or an opening up. Relatedly, Raymond Williams, in his book *Keywords*, discusses them in at least two senses: "the available and developing meanings of known words" and "the explicit but as often implicit connections which people are making" (Williams, 1976: 13). Thus behind keywords are both well-known words (elucidated by Williams's elaborations on the changing meaning of 'culture' over longer periods of time, beyond the high/low distinction) or neologistic phrases such as recent concerns surrounding 'blood minerals' or the more defused 'conflict minerals' mined and built into mobile phones. The one has readily available yet developing meanings and the other are new phraseologies that position. For the query design I am proposing, the purposive meaning of keywords is captured by Williams most readily in his second type (the new language). The first type may apply as well, such as in the case of a new use or mobilisation of a phrase, such as 'new economic order' or 'land reform'. The question then becomes what is meant by it *this time*.

Concerning how deploying a keyword implies a side-taking politics, I refer to the work of Madeleine Akrich, Bruno Latour and others, who have discussed the idea that, far from having stable meanings (as Williams also related), keywords can be parts of programs or anti-programs. Programs refer to efforts made at putting forward and promoting a particular proposal, campaign or project. Conversely, anti-programs oppose these efforts or projects through keywords. Following this reading, keywords can be thought of as furthering a program or an anti-program. There is, however, also a third type of keyword I would like to add, which refers to efforts made at being neutral. These are specific undertakings made *not* to join a program or an anti-program. News outlets such as the BBC, *New York Times* and *The Guardian* often have dedicated style guides that advise their reporters to employ particular language and avoid other. For example, the BBC instructs reporters to use generic wording for the obstacle separating Israel and the Palestinian Territories:

> The BBC uses the term 'barrier', 'separation barrier' or 'West Bank barrier' as an acceptable generic description to avoid the political connotations of 'security fence' (preferred by the Israeli government) or 'apartheid wall' (preferred by the Palestinians) (BBC Academy, 2013).

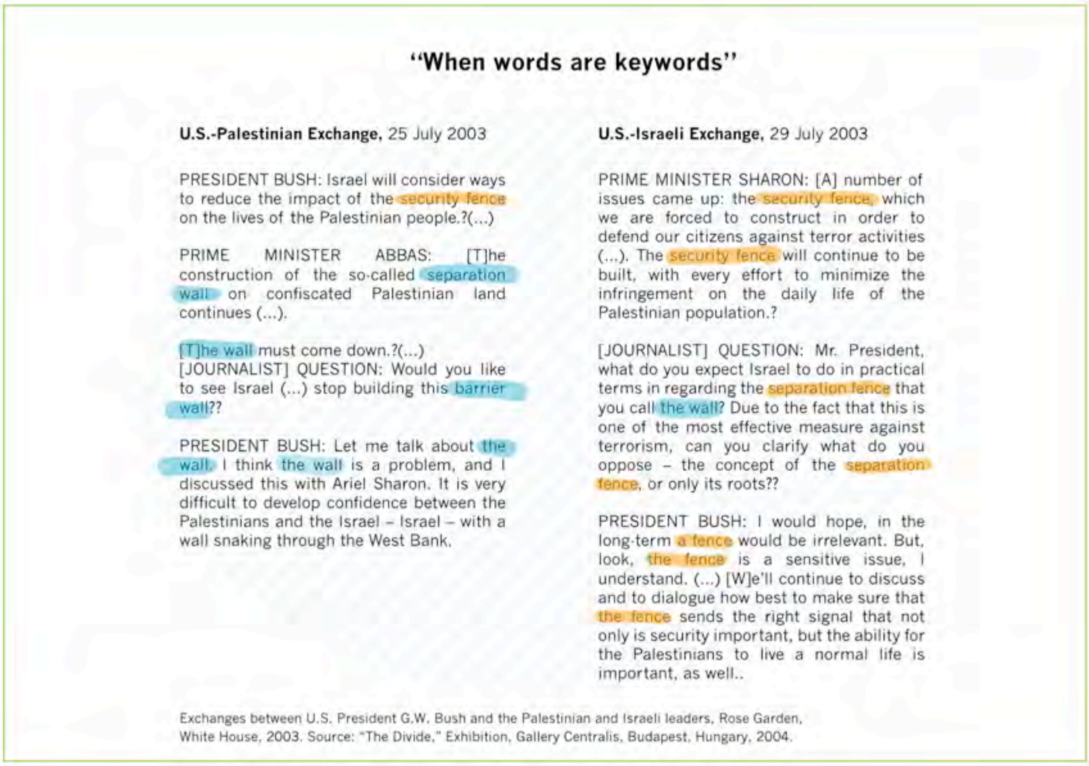
When formulating queries, it is pertinent to consider keywords as being parts of programs, anti-programs or efforts at neutrality, as this outlook allows the researcher to study trends, commitments and alignments between actors. To this end (and in contrast to discourse analysis), one does not wish to have equivalents or substitutes for the specific issue language being employed by the programs, anti-programs and the neutral programs. For example, there is a difference between using the term "blood minerals" or the term "conflict minerals", or using "blood diamonds" or "conflict diamonds", because the terms are employed (and repeated) by particular actors to issuefy, or to make into a social issue forced and often brutal mining practices that fuel war (blood diamonds or minerals) or to have industry recognise a sensitive issue and their corporate social responsibility (conflict diamonds or minerals). Therefore, they should not be treated as equivalent and grouped together. (Here it is useful to return to the point that one should use quotation marks around keywords when querying, because without quotation marks and thus specific key word queries, Google returns equivalents.) Indeed, one should treat "conflict minerals" and "blood minerals" as separate because as parts of specific programs, they show distinctive commitments and they can help to draw alignments. If someone (often a journalist) begins using a third term, such as "conflict resources", it likely constitutes a conscious effort at being neutral and not joining the programs using the other terms. Those who then enter the fray and knowledgably employ *what have become keywords* (in Williams's second sense) can be said to be taking up a position and a side, or avoiding one.

To demonstrate the notion of programs, anti-programs and efforts at neutrality further, the Palestinian-Israeli conflict, alluded to above, presents a compelling case for studying positioning as well as (temporary) alignment. There are two famous, recorded exchanges that took place at the White House in the U.S. between the then President George W. Bush and the leader of the Palestinian Authority, Mahmoud Abbas; and, secondly, between President Bush and the then Prime Minister of Israel, Ariel Sharon (see Figure 4). These exchanges, from the time when the barrier was under construction, show the kinds of positioning efforts that are made through the use of particular terms and thus the kind of specific terminology that one should be aware of when formulating queries. They also reveal temporary alignments that put on display diplomacy, with the U.S. President's using the Palestinian and then the Israeli preferred terminology in the company of the respective leaders, but only partly, thereby never fully taking sides.

The first exchange between President Bush and the Palestinian leader, Abbas, begins with a discussion where Bush refers to the barrier as a "security fence", which is the official Israeli term. Abbas then makes an attempt to correct this keyword by replying with the term "separation wall", thereby using a very different adjective – separation instead of security – to allude to the interpretation of the purpose of the barrier as separating peoples and not securing Israel. Abbas also uses a poignant noun, wall. The word "fence", as in the Israeli "security fence", connotes a lightweight, neighbourly fence. By calling it a "wall," however, Abbas connotes the Berlin Wall. The third person in this exchange, the journalist, then steps in with the term "barrier wall" in an effort not to take sides, though at the moment "wall" actually gives the Palestinian position some weight. Following this exchange, Bush, being diplomatic, realizes when talking to Abbas that the word "wall" is being used, so he switches terms and concludes by using the term, albeit without an adjective that would validate Abbas and clash with the official Israeli term.

Four days later, the Israeli Prime Minister, Sharon, visits the White House to talk to President Bush, and he begins by using "security fence", the official Israeli term. A journalist steps in and seems not to have read any newspaper style guides on the matter, because he first says "separation fence" and then "wall". The journalist, moreover, does not use "security fence" and, therefore, the question he poses, whilst critical, also seems one-sided for it was preceded by quite some Palestinian language (separation, wall). Bush concludes by being diplomatic once again to both parties involved: he is tactful to Sharon by just using the word "fence", but he does not use any adjective so as to be wary of Abbas, his recent visitor.

Wall and fence talk in the Middle East, of course, is very specific conflict terminology, but it does highlight a particular program ("security fence"), an anti-program ("separation wall") as well as an effort at being neutral ("barrier wall"). It also shows how temporary alignments, often only partial ones, are made with great tact, providing something of a performative definition of diplomacy.



"When words are keywords"

U.S.-Palestinian Exchange, 25 July 2003

PRESIDENT BUSH: Israel will consider ways to reduce the impact of the security fence on the lives of the Palestinian people.?(...)

PRIME MINISTER ABBAS: [T]he construction of the so-called separation wall on confiscated Palestinian land continues (...).

[T]he wall must come down.?(...)
[JOURNALIST] QUESTION: Would you like to see Israel (...) stop building this barrier wall??

PRESIDENT BUSH: Let me talk about the wall. I think the wall is a problem, and I discussed this with Ariel Sharon. It is very difficult to develop confidence between the Palestinians and the Israel – Israel – with a wall snaking through the West Bank.

U.S.-Israeli Exchange, 29 July 2003

PRIME MINISTER SHARON: [A] number of issues came up: the security fence, which we are forced to construct in order to defend our citizens against terror activities (...). The security fence will continue to be built, with every effort to minimize the infringement on the daily life of the Palestinian population.?

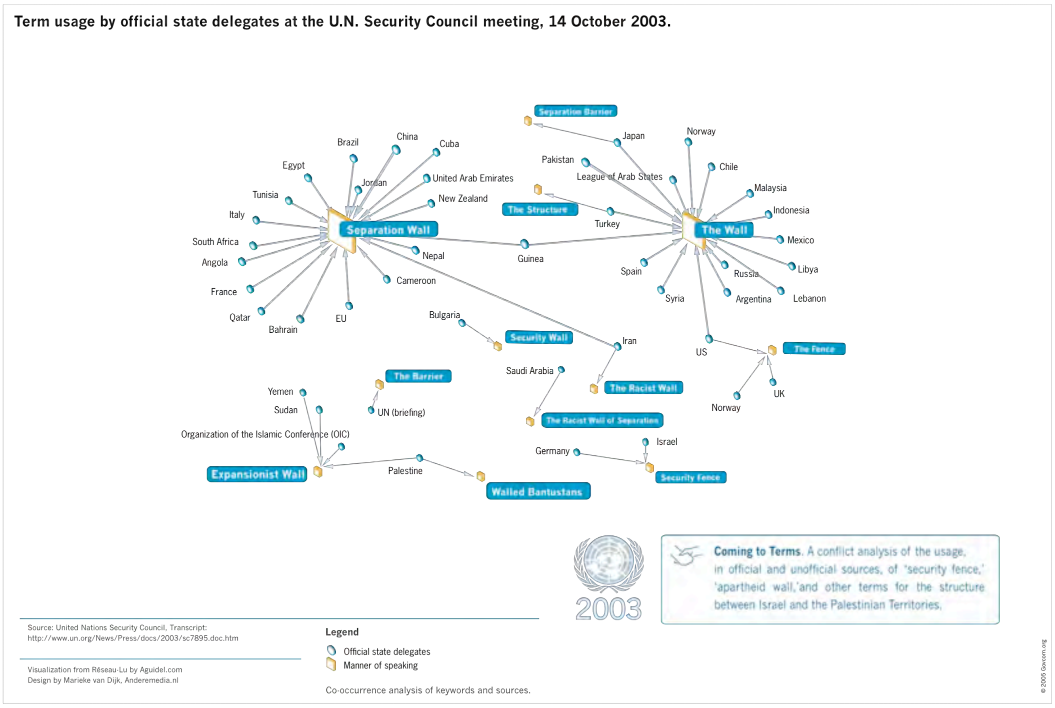
[JOURNALIST] QUESTION: Mr. President, what do you expect Israel to do in practical terms in regarding the separation fence that you call the wall? Due to the fact that this is one of the most effective measure against terrorism, can you clarify what do you oppose – the concept of the separation fence, or only its roots??

PRESIDENT BUSH: I would hope, in the long-term a fence would be irrelevant. But, look, the fence is a sensitive issue, I understand. (...) [W]e'll continue to discuss and to dialogue how best to make sure that the fence sends the right signal that not only is security important, but the ability for the Palestinians to live a normal life is important, as well..

Exchanges between U.S. President G.W. Bush and the Palestinian and Israeli leaders, Rose Garden, White House, 2003. Source: "The Divide," Exhibition, Gallery Centralis, Budapest, Hungary, 2004.

FIGURE 4. THE USE OF KEYWORDS BY U.S., PALESTINIAN AND ISRAELI LEADERS, SHOWING (TEMPORARY) TERMINOLOGICAL ALIGNMENTS AND DIPLOMACY. EXCHANGES BETWEEN THE LEADERS AT THE ROSE GARDEN, U.S. WHITE HOUSE, 2003.

Issue spaces can be analysed with this sort of keyword specificity in mind. A related example in this regard concerns the United Nations (U.N.) Security Council's debates on the barrier between Israel and the Palestinian Territories, which took place in 2003 and 2005 when it was first being constructed (Rogers and Ben-David, 2010). The terms used by each country participating in the debates were lifted directly from the Security Council transcripts. The resultant issue maps, or network graphs, contain nodes that represent countries, clustered by the term(s) that each country uses when referring to the barrier (see Figures 5 and 6). The network clearly demonstrates the specificity of the terminology put into play by the respective countries at the table as well as the terminological alignments that emerge. When countries utter the same term, groupings or blocs form, to speak in the language of international relations. For example, the largest surrounds "separation wall", and mention of other terms ("expansionist wall", "racist wall", "security wall", "the barrier", "the fence", "the wall", "the structure", "separation barrier", and so forth) make for smaller groupings or even isolation.



Term usage by official state delegates at the U.N. Security Council meeting, 14 October 2003.

Coming to Terms. A conflict analysis of the usage, in official and unofficial sources, of 'security fence,' 'apartheid wall,'and other terms for the structure between Israel and the Palestinian Territories.

Source: United Nations Security Council, Transcript: http://www.un.org/News/Press/docs/2003/sc7895.doc.htm

Visualization from Réseau-Lu by Aguidel.com
Design by Marieke van Dijk, Anderemedia.nl

Co-occurrence analysis of keywords and sources.

Figure 5. Cluster graph showing co-occurring country uses of terminology for the structure between Israel and the Palestinian Territories, UN Security Council meeting, 2003. Visualization by ReseauLu.

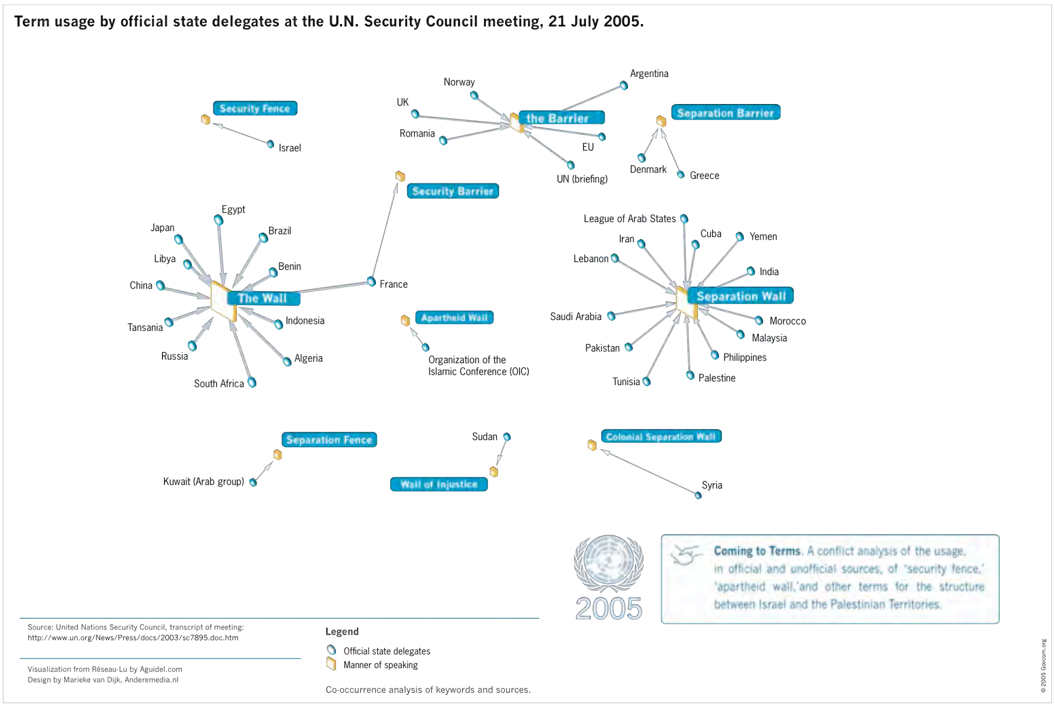

Term usage by official state delegates at the U.N. Security Council meeting, 21 July 2005.
Security Fence
Israel
Norway
UK
Romania
the Barrier
Argentina
EU
UN (briefing)
Separation Barrier
Denmark
Greece
Security Barrier
Egypt
Japan
Brazil
Libya
Benin
China
The Wall
France
Tansania
Indonesia
Russia
Algeria
South Africa
Apartheid Wall
Organization of the Islamic Conference (OIC)
League of Arab States
Iran
Cuba
Yemen
Lebanon
India
Separation Wall
Saudi Arabia
Morocco
Malaysia
Pakistan
Philippines
Tunisia
Palestine
Separation Fence
Kuwait (Arab group)
Sudan
Wall of Injustice
Colonial Separation Wall
Syria
2005
Coming to Terms. A conflict analysis of the usage, in official and unofficial sources, of 'security fence,' 'apartheid wall,'and other terms for the structure between Israel and the Palestinian Territories.
Source: United Nations Security Council, transcript of meeting: http://www.un.org/News/Press/docs/2003/sc7895.doc.htm
Visualization from Réseau-Lu by Aguidel.com
Design by Marieke van Dijk, Anderemedia.nl
Legend
Official state delegates
Manner of speaking
Co-occurrence analysis of keywords and sources.
© 2005 Govcom.org

Figure 6. Cluster graph showing co-occurring country uses of terminology for the structure between Israel and the Palestinian Territories, UN Security Council meeting, 2005. Visualization by ReseauLu.

In 2003 a majority of countries come to terms around "separation wall" or "the wall," both Palestinian side-taking terms, and there is a smattering of more extreme terms, e.g., the "racist wall". On the other side of the divide, the term "security fence", the official Israeli nomenclature, is only spoken by Israel and Germany, showing terminological alignment between the two countries. Two years later, in 2005, the next U.N. Security Council debate on the barrier took place, and a similar pattern of terminology use emerged, albeit with two distinct differences. Neutral language has found its way into the debate, with "the barrier" enjoying support. And this time, Israel is alone in using the term "security fence", and is thereby isolated.

Countries are 'linked' or isolated by terminology. They settle into a debate by subscribing to programs, anti-programs and efforts at neutrality, together with light gestures towards the one side or another (e.g., by using just wall or fence). In some cases, there are evident language blocs. Each bloc shows alignment in that countries (over time) come to terms with other countries by means of using the same language. It is precisely this alignment of actors to programs, anti-programs, or efforts of neutrality that one seeks to build into query design from the outset.

Ambiguous and unambiguous queries

If you peruses the search engine literature, there are mentions of navigational queries, transactional queries and substantive queries, among other types. Yet, on a meta level, we can broadly speak of two kinds of queries: unambiguous and ambiguous. The original strength of Google and its PageRank algorithms lay in how it dealt with an ambiguous query that matches more than one potential result and thereby is in need of some form of 'disambiguation'. An example that was often used in the early search engine literature is for the query, Harvard. This could refer to the university, a city (in Illinois, USA) or perhaps businesses near the university or in the city. By looking at which sites receive the most links from the most influential sites, PageRank would return Harvard University as the top result because it would presumably receive more links from reputable sources than a dry cleaning business near the university, for example, called Harvard Cleaners. Therefore, without unambiguous matching of keyword to result, the outputs depend on a disambiguating mechanism (Google's PageRank) that places Harvard University at the top. The ability to disambiguate is also thereby socio-epistemological or one that reveal social hierarchies. Harvard University is at the top because it has been placed there through establishment linking practices.

The social researcher may take advantage of how the search engine treats ambiguous queries. In the example, the ambiguous keyword, rights, is queried in a variety in local domain Googles (e.g., google.co.jp, google.co.uk etc.), in order to create hierarchies of concerns (rights types) per country, thereby employing Google as a socio-epistemological machine.

Contrariwise, an unambiguous query is one in which it is clear which results one is after. If we return to the cluster maps of countries using particular terms for the barrier between Israel and the Palestinian Territories, precise terms were used. By putting these terms in quotation marks and querying them, Google would return an ordered list of sources that use those specific terms. If one forgoes the use of quotation marks in the query, Google, as mentioned, ‘helpfully’ provides the engine user with synonyms or equivalents of sorts. For example, if one does not wish to make a distinction between mobile phones (British English) and cell phones (North American English), you can simply search for [mobile phones] without quotation marks and Google will furnish results for both of them. If one places a term in quotation marks, however, Google will provide results specific to that one term.

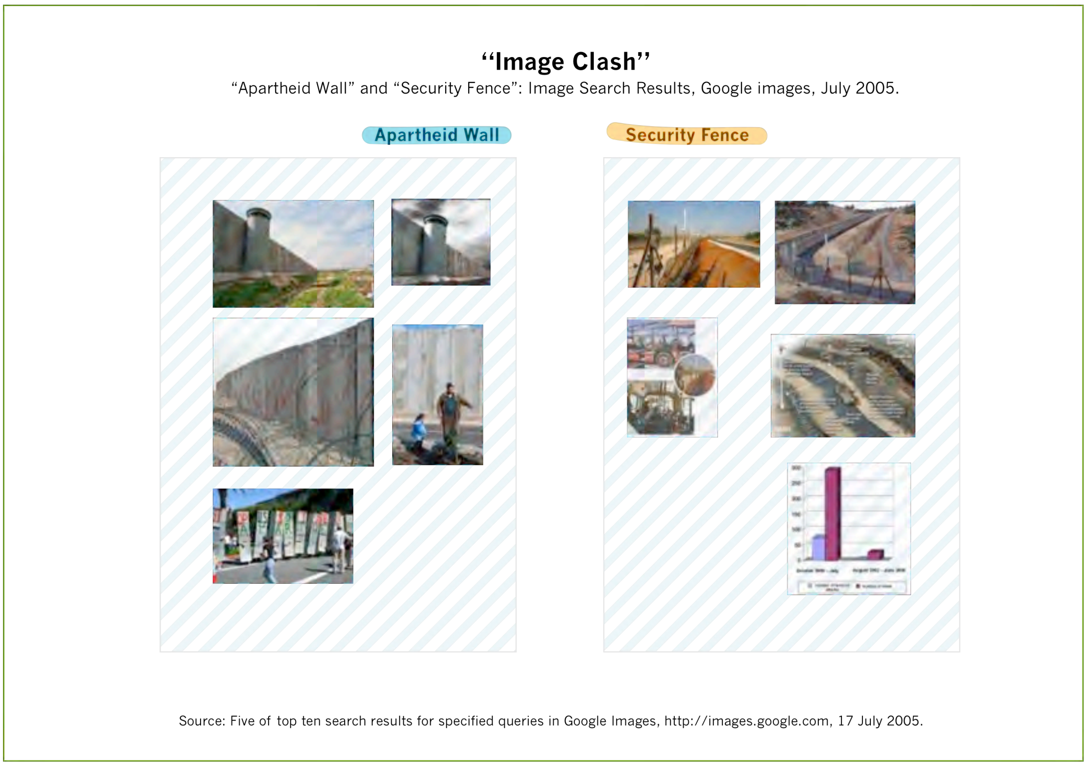
It is instructive to point out a particular form of annotation when writing about queries. When noting down the specific query used, the recommendation is to use square brackets as markers. Therefore, a query could be [“apartheid wall”], where the query has square brackets around it and the query is made as unambiguous as possible (for the engine) by using quotation marks. Oftentimes, when a query is mentioned in the literature, it will have only quotation marks without the square brackets. A reader is often left wondering whether the query was in fact made with quotation marks or whether the quotation marks are used in the text merely to distinguish the term as a query. To solve this problem, the square brackets annotation is employed. If one’s query does not have quotation marks they are dropped but the square brackets remain.

Doing search as research

There are two preparatory steps to take prior to doing search as research. The first one is to install a research browser. This means installing a separate instance of your browser, such as Firefox, or creating a new profile in which you have cleaned the cookies and otherwise disentangled yourself from Google. The second preparatory step is to take a moment to set up one’s Google result settings. If saving results for further scrutiny later (including manual interpretation as in the Rights Types project discussed below), set the results from the default 10 to 20, 50 or 100. If one is interested in researching a societal concern, one should set geography in Google to the national level – that is, to the country level setting and not to the default city setting. If one is interested in universal results only, consider obfuscating one’s location. In all cases one is not logged into Google.[4]

I would like to present, first, an example of research conducted using unambiguous queries. The project in question concerns the Google image results of the query for two different terms for the same barrier: [“apartheid wall”], which is the official Palestinian term for the Israeli-Palestinian barrier mentioned previously, versus the Israeli term, [“security fence”] (see Figure 7). The results from these two queries present images of objects distinctive from one another. The image results for [“apartheid wall”] contain graffitied, wall-like structures, barbed wire, protests, and people being somehow excluded, whereas with [“security fence”] there is another narrative, one derived through lightweight, high-tech structures. Furthermore, there is a series of images of bomb attacks in Israel, presented as justification for the building of the wall. There are also information graphics, presenting such figures as the number of attempted bombings and the number of bombings that met their targets before and after the building of the wall. In the image results we are thus presented with the argumentation behind the building of the fence. The two narratives resulting from the two separate queries are evidently at odds, and these are the sorts of findings one is able to tease out with a query design in the program/anti-program vein. Adding neutral terminology to the query design would enrich the findings by showing, for example, which side’s images (so to speak) have become the neutral ones.

When doing search as research as above, the question is often raised whether and under which circumstances to remove Google artefacts and Google properties in the results. Wikipedia, towards the top of the results for substantive queries, is ranked highly in the results for the query [“apartheid wall”] yet has as the title of its article in the English-language version an effort at neutrality in “West Bank barrier”, however much it includes a discussion of the various names given to it. Whilst a Google artefact, Wikipedia’s efforts at neutrality should be highlighted as such rather than removed. A more difficult case relates to a Google artefact in the results for an underspecified query [rights] in google.com, discussed in more detail below. The organization R.I.G.H.T.S. is returned highly in the results, owing more to its name than to its significance in the rights issue space. Here again the result was retained, and footnoted (or highlighted) as a Google artefact, which in a sense answers questions regarding the extent or breadth of artefacts in the findings. Here the research strategy is chosen to highlight rather than remove an artefact, so as to anticipate critique and make known media effects.



''Image Clash''

"Apartheid Wall" and "Security Fence": Image Search Results, Google images, July 2005.

Source: Five of top ten search results for specified queries in Google Images, http://images.google.com, 17 July 2005.

FIGURE 7. CONTRASTING IMAGES FOR ["APARTHEID WALL"] AND ["SECURITY FENCE"] IN GOOGLE IMAGES QUERY RESULTS, JULY 2005.

As the last example, I would like to present a project using an ambiguous query that takes advantage of Google's social sorting. In this case we undertook a project about rights, conducted by a large group of researchers who spoke some 30 languages amongst them. Using this abundance of diverse language skill, we set about to determine which sorts of rights are held dear to particular cultures relative to others. In the local languages we formulated the query for [rights], and we ran the query in all the various local domain Googles per language spoken, interpreting the results from google.se as Swedish concerns, .fi for Finnish, .ee for Estonian, .lv for Latvian, .co.uk for British, and so forth. With the results pages saved as HTML (for others to check), the researchers were instructed to work with an editorial process where they manually extract the first 10 unique rights from the search results of each local domain Google.[5]

Information designers visualized the results by creating an icon for each right type and a colour scheme whereby unique rights and shared rights across the languages were differentiated. The resultant infographic graphically shows rights hierarchies per country as well as those rights that are unique to a country and those shared amongst two or more countries. One example of a unique right is the case of Finland, in which the "freedom to roam" is high on the list (see Figure 8). Far from being a trivial issue, what this freedom means is that one can walk through someone's backyard, whereas in other countries (e.g., the U.K.), it is not a right, and organizations lobbying for the right to ramble and walk the ancient pathways. Another example is in Latvia, where pension rights for non-citizens are of particular importance.

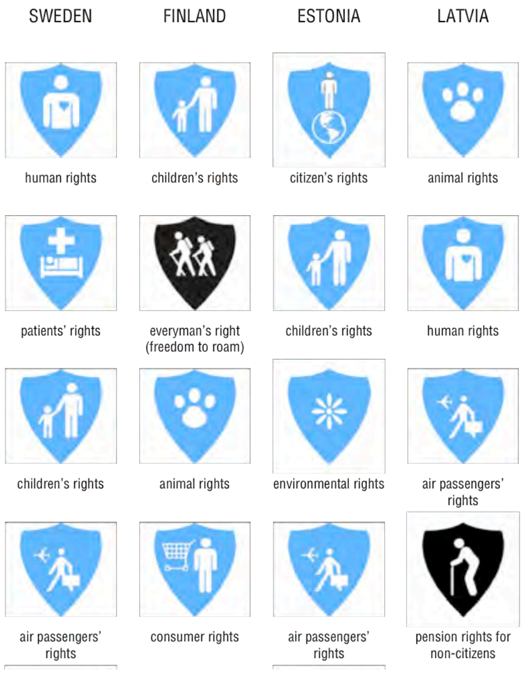


FIGURE 8. RIGHTS TYPES IN PARTICULAR COUNTRIES, RANKED FROM GOOGLE RESULTS OF THE QUERY [RIGHTS] IN THE LOCAL LANGUAGES AND LOCAL DOMAIN NAME GOOGLES (GOOGLE.SE, GOOGLE.FI, GOOGLE.EE AND GOOGLE.LT), JULY 2009.

Conclusions

Digital methods have been developed as a distinctive strategy for Internet-related research where the web is considered an object the study for more than online or digital culture only. As a part of the computational turn in social research, digital methods were developed as a counterpart to virtual methods, or the importation of the social scientific instrumentarium into the web, such as online surveys. Digital methods, as an alternative, strive to employ the methods of the medium, imagining the research affordances of engines and platforms, and repurposing their methods and outputs for social (and medium) research.

The contribution here is foundational is the sense of outlining certain premises of digital methods but also the nitty-gritty of doing online analysis. In conclusion, I would like to return to the premises of doing digital methods with Google Web Search in particular as well as to the finer points of query design, which underpins 'search as research' as an approach distinctive to other analytical traditions, such as discourse and content analysis.

First, in the digital method, search as research, Google is repurposed from its increasing use as a consumer information appliance, with personalised results that evermore seek to anticipate consumer information needs (such as with autosuggest as well as the service, Google Instant). Rather, Google is relied upon as an epistemological machine, yielding source hierarchies and dominant voice studies (through its ranked results for a keyword query) as well as individual actor commitment (through its quantitative counts for a single or multiple site query). Transforming Google back into a research machine (as its founders asserted in the early papers on its algorithms) these days requires disentangling oneself from the engine through the installation of a clean research browser and logging out. Once in use, the research browser is not expected to remove all Google artefacts from the output (e.g., Google properties, SEO'd results, etc.), but in the event they become less obfuscated and an object of further scrutiny (medium research) together with the social research one is undertaking with repurposed online methods.

Query design is the practice behind search as research. One formulates queries whose results will allow for the study of trends, dominant voice, positioning, commitment, concern and alignment. The technique is sensitive to keywords, which are understood as the connections people are currently making of a word or phrase, whether established or neologistic, leaning on Raymond Williams second definition of a keyword. Indeed, in the query design put forward above, the keywords used could be said to take sides, and are furthermore conceptualised as forming part of a program or anti-program, as developed by

Madeleine Akrich and Bruno Latour. I have added a third means by which keywords are put into play. Journalists, and others conspicuously not taking sides, develop and employ terms as efforts at neutrality. [“West Bank barrier”] is one termed preferred by BBC journalists (and the English-language Wikipedia) over [“security fence”] (Israeli) or [“apartheid wall”]. Querying a set of sources (e.g., country speeches at the U.N. Security Council debates) for each of the terms and noting use as well as common use (co-occurrence) would show positioning and alignment, respectively.

Secondly, for digital methods practice, I would like to emphasize that for query design in the conceptual framework of program/anti-program/efforts at neutrality, one retains the specific language (instead of grouping terms together), because the exact matches are likely to show alignment and non-alignment. Furthermore, language may also change over time. Therefore, if one conducts an overtime analysis, one can determine whether or not certain actors have, for example, left a certain program and joined an anti-program by changing the language and terms they use. Some countries may have become neutral, as was noted when contrasting term use in the 2003 versus the 2005 Security Council debates on the barrier. As another example, one could ask, has there been an alignment shift signified through actors leaving the “blood minerals” program and joining the “conflict minerals” program?

Thirdly, whilst the discussion has focused mainly on unambiguous queries, search as research also may take advantage of ambiguous ones. As has been noted, if we are interested researching dominant voice, commitment and showing alignment and non-alignment, an unambiguous query is in order. Through an ambiguous query, such as [rights], one can tease out differences and distinct hierarchies of societal concerns across cultures. Here a cross-cultural approach is taken which for search as research with Google implies a comparison of the results of the same query (albeit in each of the native languages) of local domain Google results.

Finally, query design may be viewed as an alternative to forms of discourse and content analysis that construct labelled category bins and toss keywords (and associated items) into them. That is, in query design specificity of the language matters for it differentiates as opposed to groups. More generally, it allows one to cast an eye onto the entire data set, making as a part of the analysis so-called long tail entities that previously would not have made the threshold. One studies it all without categorising and without sampling, which (following Akrich and Latour), allows not only for the actors to speak for themselves and for the purposes of their program, anti-program or efforts at neutrality, but (following Lev Manovich's cultural analytics) provides opportunities for new interpretive strategies. That there arises a new hermeneutics (that combines close and distant reading) could also be seen as the work ahead for the analytical approach.[6]

Notes

[1] The U.S. Center for Disease Control and Prevention ran a competition in 2013-2014 for instruments that use search and social media data to forecast influenza, and the one employing the data from Google Flu Trends won the award.

[2] Crowdfindings is a term coined by Christian Bröer.

[3] Not only Google Trends but also Google Related Search provide means for studying keyword salience as well as the association between keywords, including co-occurrence.

[4] It is also important to note that simply using private browsing tools, such as the incognito tool on Google Chrome, does not suffice as a disentanglement strategy for that only prevents the saving of one's search history to one's machine. It is still being saved at headquarters so to speak. When in incognito mode, one is still served personalized results.

[5] According to Google's terms of service, one is not allowed to save results, or make derivative works from them. The research thus could be considered to break the terms of service, however much the spirt of those terms is to prevent commercial gain through redistribution rather than to thwart academic research. The results pages are saved as HTML, with a uniform naming convention so that one could return to them, and, in recognition of the terms of service, were not shared to a data repository.

[6] At the lecture delivered at the Digital Methods Winter School, January 2015, Lev Manovich proposed work on a 'new hermeneutics' after the study and visualisation of 'all data', substituting continuous change for periodization and and continuous description for categorisation.

Acknowledgments

The author would like to thank Anat Ben-David for working on and giving presentations about query design that led to this article and Becky Cachia for editorial assistance. Michael Stevenson, Erik Borra and researchers at the Digital Methods Initiative, University of Amsterdam, provided crucial input.

References

Akrich, Madeleine and Bruno Latour (1992). "The De-Scription of Technical Objects," in Wiebe Bijker and John Law (eds.), *Shaping Technology / Building Society: Studies in Sociotechnical Change*. Cambridge, MA: MIT Press, 205-224.

BBC Academy (2013). "Israel and the Palestinians," Journalism Subject Guide. London: BBC, http://www.bbc.co.uk/academy/journalism/article/art20130702112133696.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski and Larry Brilliant (2009). "Detecting influenza epidemics using search engine query data," *Nature* 457, 1012-1014.

Hindman, Matthew (2008), *The Myth of Digital Democracy*. Princeton: Princeton University Press.

Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani (2014). "The Parable of Google Flu: Traps in Big Data," *Science* 343:6176, 1203-1205.

Pariser, Eli (2011). *The Filter Bubble*. New York: Penguin Press.

Negroponte, Nicholas (1995). *Being Digital*. London: Hodder and Stoughton.

Rogers, Richard and Anat Ben-David (2010). "Coming to Terms: A conflict analysis of the usage, in official and unofficial sources, of 'security fence,' 'apartheid wall,' and other terms for the structure between Israel and the Palestinian Territories," *Media, Conflict & War*. 2(3): 202-229.

Sunstein, Cass (2001). *Republic.com*. Princeton: Princeton University Press.

Turow, Joseph (2006). *Niche Envy*. Cambridge, MA: MIT Press.

U.S. Center for Disease Control (2014). “CDC Announces Winner of the ‘Predict the Influenza Season Challenge’”, Press release, 18 June, http://www.cdc.gov/flu/news/predict-flu-challenge-winner.htm.

Williams, Raymond (1975). *Keywords: A Vocabulary of Culture and Society.* London: Fontana.

◆ SPECIAL FEATURES ◆

POLITICAL RESEARCH IN THE DIGITAL AGE

by Richard Rogers

Professor of New Media and Digital Culture, Department Chair of Media Studies at the University of Amsterdam and Director of the Digital Methods Initiative

INTRODUCTION: COMPUTATIONAL OR DIGITAL TURN?

There is currently a debate at hand over aligning political and social research with the digital age (boyd and Crawford 2012). How to cope with the challenges the Internet and the digital, including newly available online data, bring to research? Concomitant with the rise of the term Big Data, certain methods and tools appear to drive research as well as the complex of what could be called the programmatic agenda, e.g., special issues of journals, funding calls, conference titles, lecture series and so forth. For some, it has been termed the computational turn, meaning the importation of computer science techniques into social research practices (Berry 2011). More dramatically, that turn supposedly comes with paradigm-rending consequences such as pattern-seeking supplanting interpretation (Savage and Burrows 2007; Watts 2007; Lazer et al. 2009). Another, subtly different means of phrasing the arrival of the stickered laptops and hacking workshop culture could be the digital turn, where the study of digital culture informs research that makes use of online data, software and visualizations. To make this distinction between the computational and the digital turns is also a means of resisting a monolithic, or unitary, understanding of the changing nature of research in the digital age (Lovink, 2014). More specifically, there are variegated approaches across the digital humanities, e-social sciences as well as digital media studies that could be seen as having distinctive ontological and epistemological commitments and positionings. Here I briefly situate and discuss a series of digital research practices called cultural analytics, culturomics, webometrics, altmetrics and digital methods, providing short examples of what they could offer in terms of political research (Manovich 2011; Michel et al. 2010; Priem et al. 2010; Rogers 2013). First, each may be differentiated according to their preferred materials as well as methodological outlook, which I have previously described in terms of working with the digitised (materials and methods), the natively digital or some combination (see also Rogers 2009). Second, instead of translating political research practices for the web (e.g., searching for the public sphere in forums, striving to locate public debate in the comment space or undertaking online surveying and polling), the invitation issued by the digital turn is more experimental, and perhaps interdisciplinary. How to repurpose the computational and digital techniques for political studies? Finally, I concentrate on a new space for political expression (Facebook), and briefly put forward an analytics approach to studying engagement, a typical concern in political research that is operationalized as a digital method combining counting and interpretation.

DIGITISED, NATIVELY DIGITAL OR SOME COMBINATION

To begin, an ontological distinction may be made between the materials "of the medium" and those that have migrated to it (Blood 2007). Blogs, considered of the web, are in this rendering natively digital, whereas a scanned book, made available through Google Books, is a digital newcomer, or digitised material. Another conceptual means of making the distinction are webpages that cannot be printed, but rather screen-grabbed only (Latour 2004). The distinction between the natively digital and the digitised also may be applied to methods. There are those methods that have been migrated to the web, such as online surveys, and those written for it, such as Google's PageRank (privileging one website over another in a ranking) or Facebook's EdgeRank (privileging friends over others in terms of closeness). Approaches in digital research thus may be arrayed in terms of which materials are the preferred data (digitised or born-digital) and where the methods are situated (emulated or native) (see table below).

Table One: Situating five approaches to digital humanities and e-social sciences according to their preferred data and method types.

		METHOD	
		DIGITISED	NATIVELY DIGITAL
DATA	DIGITISED	► Culturomics* ► Cultural Analytics*	► Altmetrics
	NATIVELY DIGITAL	► Webometrics	► Digital Methods

DIGITAL RESEARCH FIVE WAYS

Over the past decade the methods and techniques developed for digital research (using both digitised as well as online data) have been couched in a variety of descriptors, with notions of analytics, metrics, -nomics or methods appended, providing rather different emphases in what is being measured. Analytics is most closely associated with the platform industries (Google, YouTube, Facebook, Twitter, Pinterest, Adobe and others), connoting pattern recognition in (user) data. One captures and analyses user (interaction) data, populating dashboards and other interfaces with visualizations aiming to provide "actionable insights," as the software company Adobe phrases it (Adobe 2014). Metrics are standards of measurement and take their nomenclature from counting techniques in library and information science, including bibliometrics and scientometrics. One is concerned with such measures as impact, salience, and resonance, meaning not only the brute force, but its relative strength and endurance. The choice of the suffix -nomics is perhaps furthest from online industry-science relations, and refers to law, as in the laws of nature, connoting fundamental discovery or basic pursuits. It has in common with the term "methods" a more open-ended epistemology. However one goes about the study, and with whichever approach, methods emphasize a procedure or research protocol with steps. When described as such, digital methods could cover the range of procedures to study digital materials, not merely online methods for studying web data, as I come to after a brief discussion of cultural analytics, culturomics, webometrics and altmetrics, providing means to rework each for political research.

Cultural analytics, the first of the named approaches in digital humanities, often uses as its materials digitised collections, such as the covers of a tone-setting magazine like *Time* or the oeuvre of an artist. It has a preferred piece of software, ImagePlot, which groups images according to formal properties, including hue and saturation. It may be used to make chronologies, such as of the images made of the Gezi Park protests in Istanbul in May and June 2013. Using the technique, one notes the transformation of Turkey's so-called "tree revolution," where, as one eyewitness explained it, "the conversion of public space into private space explain[s] why the occupation of Gezi Park is not just meant to save trees, but to save Turkey's democracy" (see Figure One ong page 81.) (Turkey EJOLT Team 2013). Green imagery gradually declines, yielding to images of protesters being pepper-sprayed and more generally to rights fights.

Culturomics, a second digital humanities approach, queries Google's collection of digitised books (via the Google Ngram Viewer) for words, thereby displaying cultural or societal trends, most robustly from English-language books published between 1800 and 2000, though there are collections of books from other languages, too. The outputs are keyword graphs, showing frequency of mentions over time. In technique and visual style, the graphing echoes the earlier Google Insights tool, which showed the incidence of keywords users sought in search queries. Searches may be political, for particular queries may land on right-leaning or left-leaning websites. For example, in the run-up to the American presidential elections in 2012, users who queried for "obamacare" landed predominantly on right-leaning websites, and for "obama student loan forgiveness" on left-leaning sites (see Figure Two on page 83) (Borra and Weber 2012). Keyword query analysis may also include users' geolocation, thus inviting work on the use of terms by geography. One could consider geolocating hate speech (via queries for particular language) and observing its steadiness or fluctuation longitudinally.

In the e-social sciences, webometrics are citation analysis methods using web links (mainly) as if they were academic citations, where a link is treated as an endorsement or impact metric (Thelwall et al. 2005). Webometric approaches are built into software such as IssueCrawler and VOSON that crawl websites, locate linking and visualize relationships as network graphs, thereby showing the characteristics of the network, including the centrality or peripherality of one or more specific actors. It may also show an online strategy, as depicted in the IssueCrawler network graphs made of Barack Obama's online campaign in 2008 (Venturini 2010). The exceptional star shape of the network is caused by the campaign's strategy of linking (see Figure Three on page 84). The core of the network is formed by barackobama.com and its subsites, such as latinos.barackobama.com, faith.barackobama.com and students.barackobama.com. The periphery consists mainly of social media sites about Obama, and features his pages on LinkedIn, Facebook, Flickr, etc. The network also crowds out other websites, thereby displaying not the grassroots, new media campaigning style employed by Howard Dean in 2004 (which allowed users to create their own narratives during sponsored meet-ups), but rather a stay-on-message approach (Rogers 2005).

Altmetrics inverts traditional scientometrics, counting citations of academic work that appear not in published journals, but rather in blogs, on Twitter or in other online spaces. Counting (and interpreting) references in social media is part of a larger analytical approach to the substance and source commitments of a topical, issue or ideological network, e.g., on Facebook or Twitter. For example, one may note the top referenced content (in this case most linked-to webpages) by Ministry-level Dutch civil servants on Twitter. It was found that civil servants tend to follow news, politicians and new media and political trend-watchers, as opposed to citizens, who are absent (see Figure Four on page 85). The work that is most referenced, moreover, concerns civil servant use of new media as well as innovative online campaigns and initiatives, meaning the content shared is self-referential and medium-related, in the first instance, rather than otherwise topical.

As mentioned above, some may employ the term digital methods to cover the entirety of the digital turn techniques described above, or, increasingly, "mainstream" research techniques (Venturini 2010). More specifically, it refers to repurposing online devices and platforms (such as Google searches, Facebook and Wikipedia) for social and political research that would often have been otherwise improbable. Among the tools developed is the so-called Lippmannian device, a Google Scraper that detects bias or leaning of an actor on the basis of the type of keyword mentions (see Figure Five on page 86). Thus one may query a set of climate change websites for mention of the names of climate change skeptics, thereby finding skeptic-friendly actors (as well as watchdog sites that also follow and mention them). In the above case, Google is repurposed as a research machine rather than its typical use as a consumer information appliance.

DIGITAL METHODS ENCOURAGE A SOCIOLOGICAL OUTLOOK OR IMAGINATION ABOUT RESEARCH OPPORTUNITIES THAT EXIST IN ONLINE CULTURE.

CONCLUSION: FOLLOWING THE MEDIUM AS A STARTING POINT FOR DIGITAL RESEARCH

Digital Methods, either generally or more specifically as the practice of repurposing devices, are not just toolkits or operating instructions for software packages; they deal with broader questions about how to do research online. They encourage a sociological outlook or imagination about research opportunities that exist in online culture by following the medium rather than asking it to do one's disciplinary bidding. One case in point, by way of conclusion, is the study of political activism. One could critique the rise of slacktivism or clicktivism, online activities that require little in the way of commitment but give one the feeling of having done something for the cause. Alternatively, one might study how liking, sharing and commenting on particular content show engagement, thereby studying (for instance) which videos or photos are currently animating anti-Islam groups and pages in Facebook (see Figure Six on page 87). The study of engagement borrows here from an analytics framework that captures clicks as well as comments, and identifies the content that animates, opening up opportunities for further interpretation. Here the call is to rely at the outset on medium activity measures and ask what might be learned from them.

SOURCES

Ackland, Robert. 2013. *Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age*. London: Sage.

Adobe. 2014. "Adobe Analytics." last modified May 2, 2014, http://www.adobe.com/solutions/digital-analytics.html, accessed May 2, 2014.

Baetens, Tom, Tom Juetten, Jan Maessen, Erik Borra, and Richard Rogers. 2013. *De Uitzondering op de Regel: Over Ambtenaren in de Openbaarheid*. The Hague: Ministry of Internal Affairs / Emma Communicatie.

Berry, David M. 2011. "The Computational Turn: Thinking About the Digital Humanities." *Culture Machine* 12: 1-22.

Blood, Rebecca. 2002. "Introduction." In *We've Got Blog: How Weblogs Are Changing Our Culture*, edited by John Rodzvilla, ix-xiii. Cambridge, MA: Perseus.

Borra, Erik and Ingmar Weber. 2012. "Political Insights: Exploring Partisanship in Web Search Queries." *First Monday* 17. http://firstmonday.org/ojs/index.php/fm/article/view/4070/3272, accessed May 2, 2014.

boyd, danah and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15: 662-679.

Foot, Kirsten A. and Steven M. Schneider. 2006. *Web Campaigning*. Cambridge, MA: MIT Press.

Krippendorff, Klaus H. 2012. *Content Analysis: An Introduction to its Methodology*. London: Sage.

Latour, Bruno. 2004. "Paris: Invisible City." http://www.bruno-latour.fr/virtual/EN/index.html, accessed May 2, 2014.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Sinan, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy and Marshall Van Alstyne. 2009. "Computational Social Science." *Science* 323: 721-723.

Manovich, Lev. 2011. "Trending: The Promises and the Challenges of Big Social Data." http://www.manovich.net/DOCS/Manovich_trending_paper.pdf, accessed May 2, 2014.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak and Erez Lieberman Aiden. 2011. "Quantitative Analysis of Culture Using Millions of Digitised Books." *Science* 331: 176-182.

Priem, Jason, Dario Taraborelli, Paul Groth and Cameron Neylon. 2010. "alt-metrics: a manifesto." http://altmetrics.org/manifesto/, accessed May 2, 2014.

Rogers, Richard. 2005. "Old and New Media: Competition and Political Space." *Theory and Event* 8 https://muse.jhu.edu/journals/theory_and_event/v008/8.2rogers.html, accessed May 2, 2014.

Rogers, Richard. 2009. *The End of the Virtual*. Amsterdam: Amsterdam University Press.

Rogers, Richard. 2013. *Digital Methods*. Cambridge, MA: MIT Press.

Savage, Mike and Roger Burrows. 2007. "The Coming Crisis of Empirical Sociology." *Sociology* 41: 885-899.

Thelwall, Mike, Laura Vaughan and Lennart Björneborn. 2005. "Webometrics," in *Annual Review of Information Science and Technology* 39, edited by Blaise Cronin, 81-135. Medford, NJ: Information Today.

Turkey EJOLT Team. 2013. "Turkey's Tree Revolution – part 2: Everyday I'm chapulling" *EJOLT: Mapping Environmental Justice*. http://www.ejolt.org/2013/06/turkeys-tree-revolution-part-2-everyday-im-chapulling/, accessed May 2, 2014.

Venturini, Tommaso. 2012. "Building on Faults: How to Represent Controversies with Digital Methods." *Public Understanding of Science* 21: 796–812.

Watts, Duncan J. 2007. "A Twenty-first Century Science." *Nature* 445: 489.

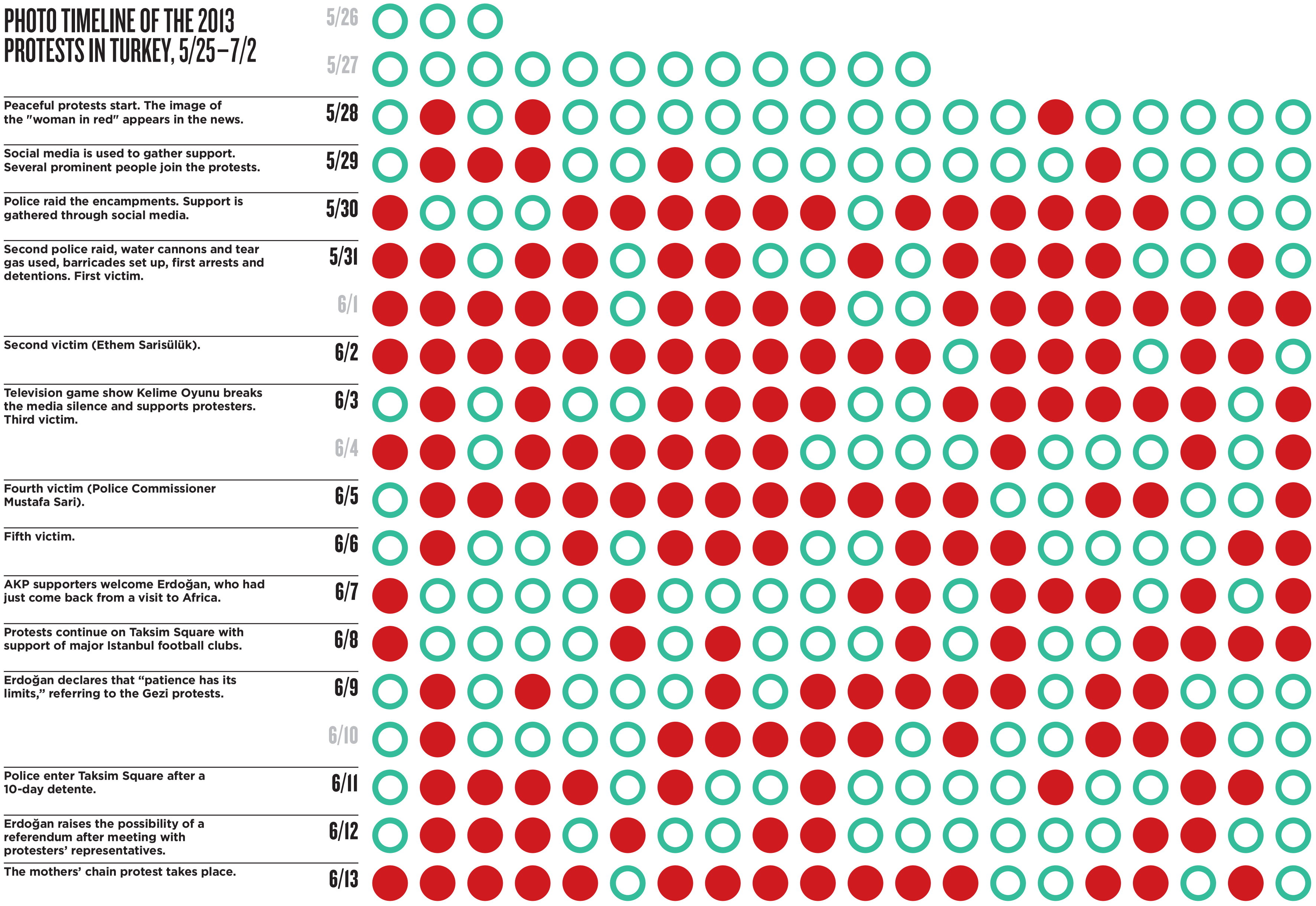
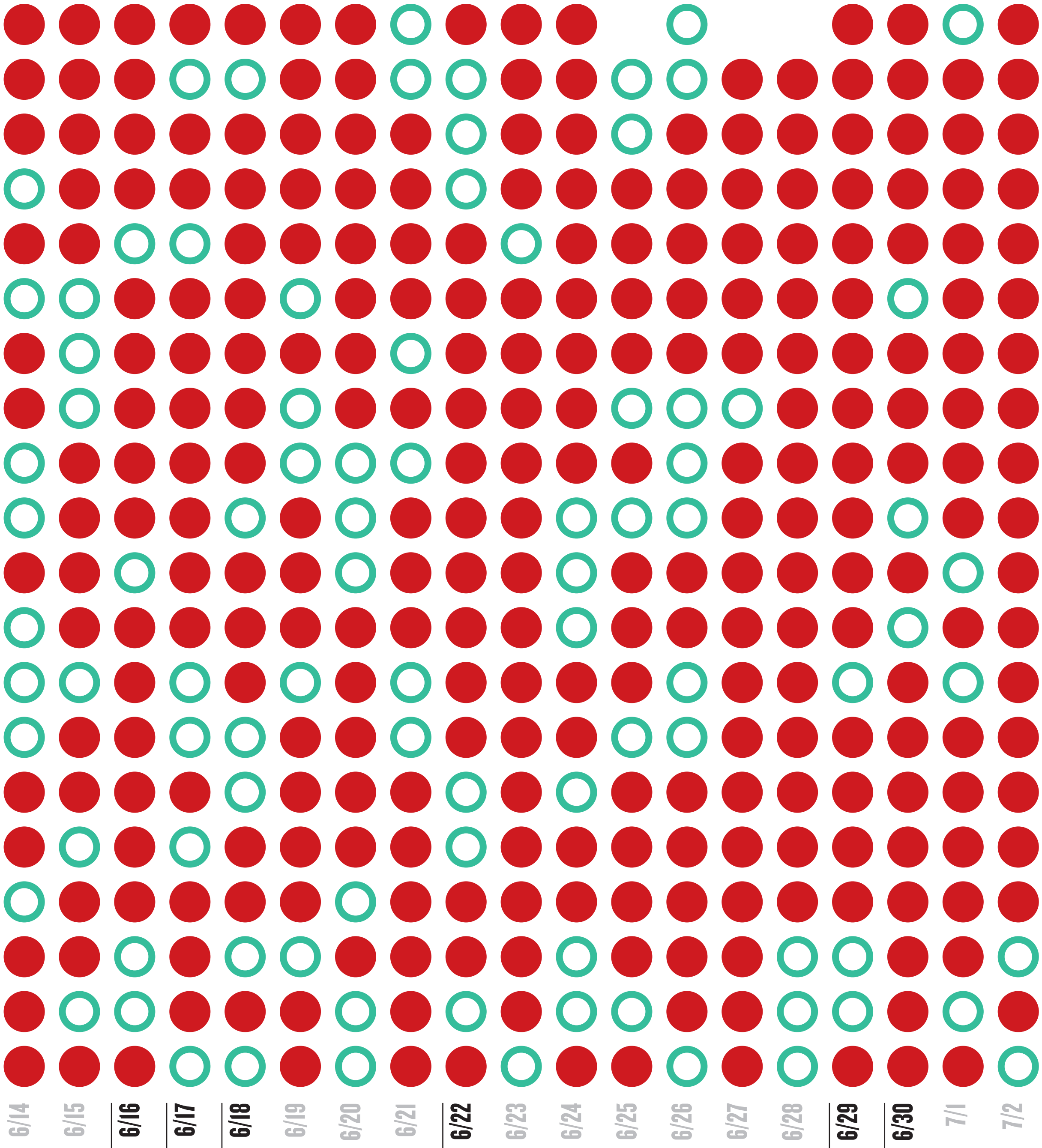


Figure One: Image characterization of top images returned from Google Images, query [Gezi] according to "save the trees" (green outlines) or "bring down the government" (red fills), June 2013. (cc) Digital Methods Initiative, Amsterdam, 2013.



6/16 Protests and police violence continue.

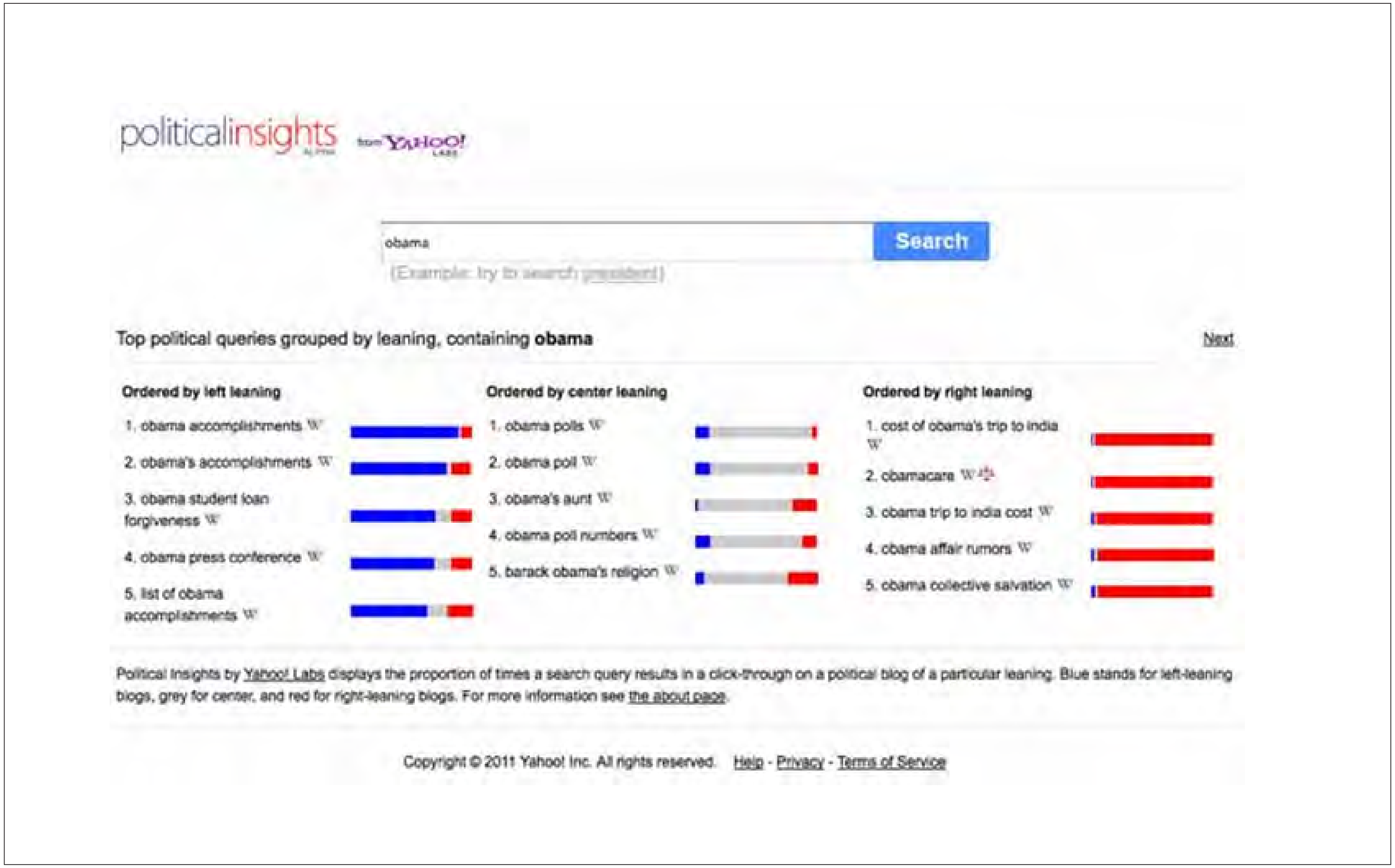
6/17 Protest grows, becomes national.

6/18 Silent protest imitates and supports Erdem Gündüz, the "standing man."

6/22 Protesters throw flowers at the police; police attacks. Mass demonstrations around Turkey.

6/29 Demonstrations in Taksim Square and in Ankara against the release of police officer Ahnet Sahbaz. Suppression of protesters.

6/30 Massive march, football team fans join the protest.



Top political queries grouped by leaning, containing **obama**

Ordered by left leaning

1. obama accomplishments W
2. obama's accomplishments W
3. obama student loan forgiveness W
4. obama press conference W
5. list of obama accomplishments W

Ordered by center leaning

1. obama polls W
2. obama poll W
3. obama's aunt W
4. obama poll numbers W
5. barack obama's religion W

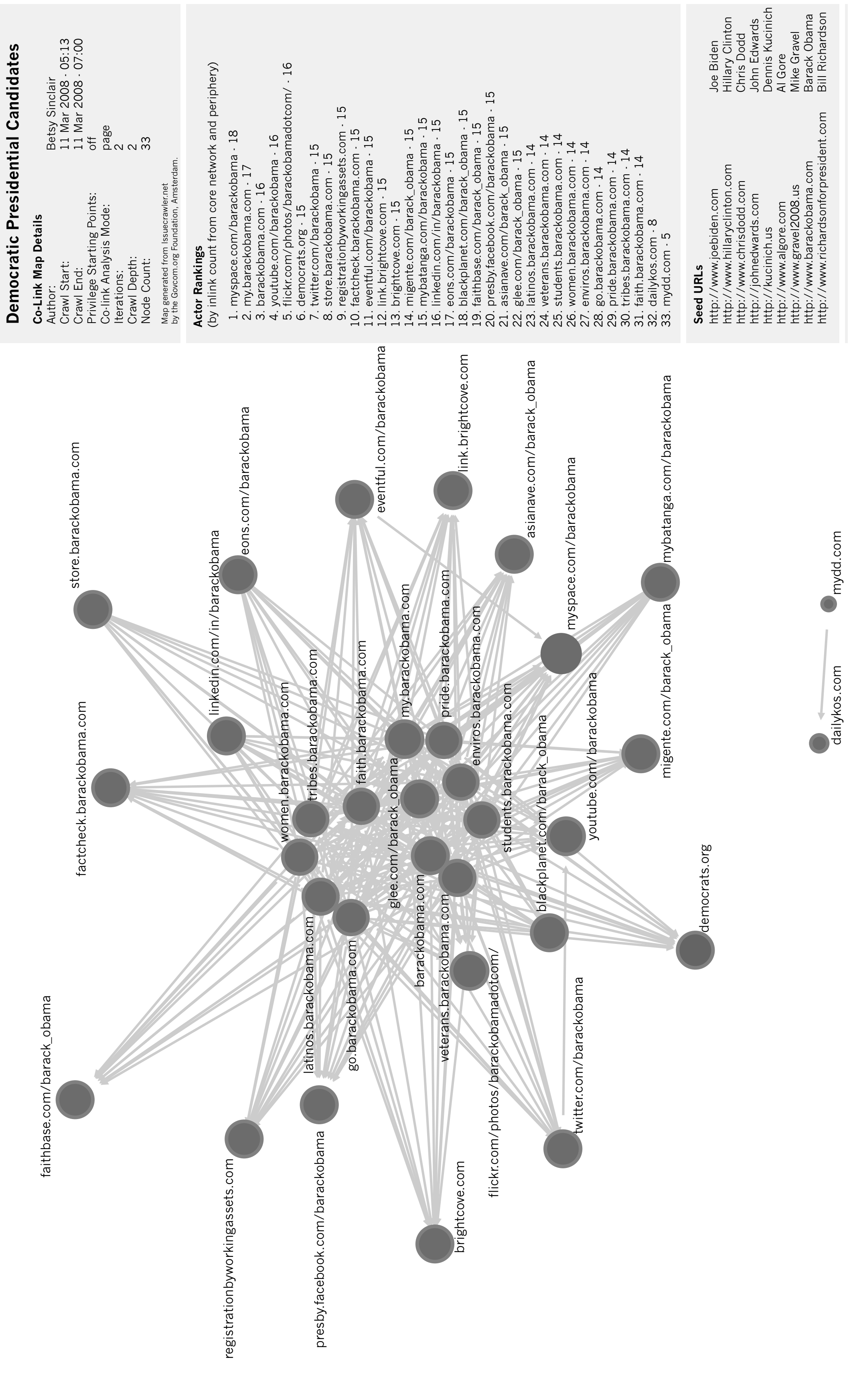
Ordered by right leaning

1. cost of obama's trip to india W
2. obamacare W
3. obama trip to india cost W
4. obama affair rumors W
5. obama collective salvation W

Political Insights by Yahoo! Labs displays the proportion of times a search query results in a click-through on a political blog of a particular leaning. Blue stands for left-leaning blogs, grey for center, and red for right-leaning blogs. For more information see the about page.

Copyright © 2011 Yahoo! Inc. All rights reserved. Help - Privacy - Terms of Service

Figure Two: Political Insights, Yahoo! Labs, showing right-leaning and left-leaning queries related to Obama, 2011. Source: Borra and Weber, 2012.



Democratic Presidential Candidates

Co-Link Map Details

Author:	Betsy Sinclair
Crawl Start:	11 Mar 2008 · 05:13
Crawl End:	11 Mar 2008 · 07:00
Privilege Starting Points:	off
Co-link Analysis Mode:	page
Iterations:	2
Crawl Depth:	2
Node Count:	33

Map generated from Issuecrawler.net
by the Govcom.org Foundation, Amsterdam.

Actor Rankings
(by inlink count from core network and periphery)

1. myspace.com/barackobama · 18
2. my.barackobama.com · 17
3. barackobama.com · 16
4. youtube.com/barackobama · 16
5. flickr.com/photos/barackobamadotcom/ · 16
6. democrats.org · 15
7. twitter.com/barackobama · 15
8. store.barackobama.com · 15
9. registrationbyworkingassets.com · 15
10. factcheck.barackobama.com · 15
11. eventful.com/barackobama · 15
12. link.brightcove.com · 15
13. brightcove.com · 15
14. migente.com/barack_obama · 15
15. mybatanga.com/barackobama · 15
16. linkedin.com/in/barackobama · 15
17. eons.com/barackobama · 15
18. blackplanet.com/barack_obama · 15
19. faithbase.com/barack_obama · 15
20. presby.facebook.com/barackobama · 15
21. asianave.com/barack_obama · 15
22. glee.com/barack_obama · 15
23. latinos.barackobama.com · 14
24. veterans.barackobama.com · 14
25. students.barackobama.com · 14
26. women.barackobama.com · 14
27. enviros.barackobama.com · 14
28. go.barackobama.com · 14
29. pride.barackobama.com · 14
30. tribes.barackobama.com · 14
31. faith.barackobama.com · 14
32. dailykos.com · 8
33. mydd.com · 5

Seed URLs

http://www.joebiden.com	Joe Biden
http://www.hillaryclinton.com	Hillary Clinton
http://www.chrisdodd.com	Chris Dodd
http://johnedwards.com	John Edwards
http://kucinich.us	Dennis Kucinich
http://www.algore.com	Al Gore
http://www.gravel2008.us	Mike Gravel
http://www.barackobama.com	Barack Obama
http://www.richardsonforpresident.com	Bill Richardson

Figure Three: Issuecrawler graph of interlinking among Obama-related websites, 2008.
Source: Issuecrawler.net, © Govcom.org Foundation, 2008, published in Krippendorff, 2012.

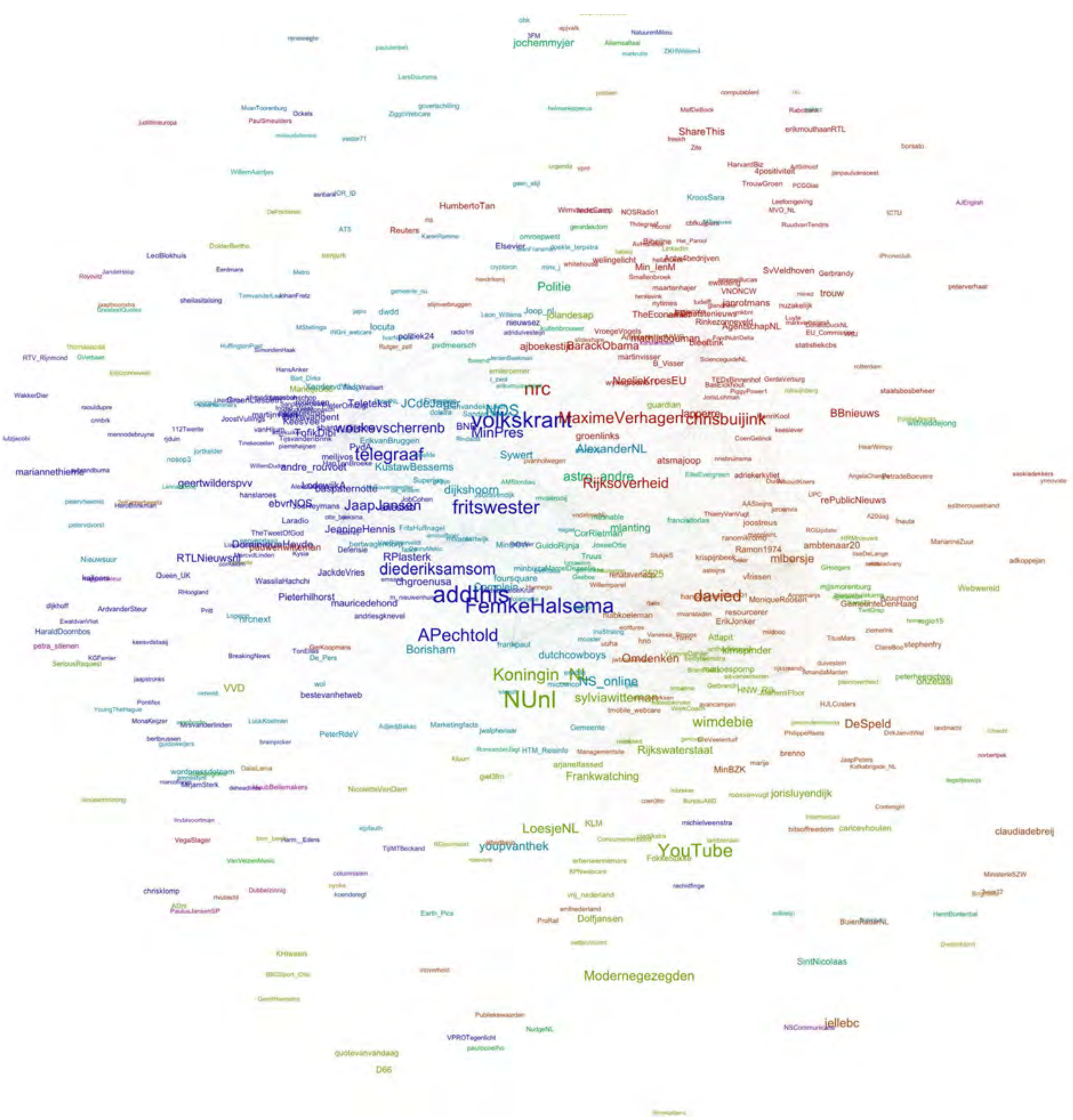
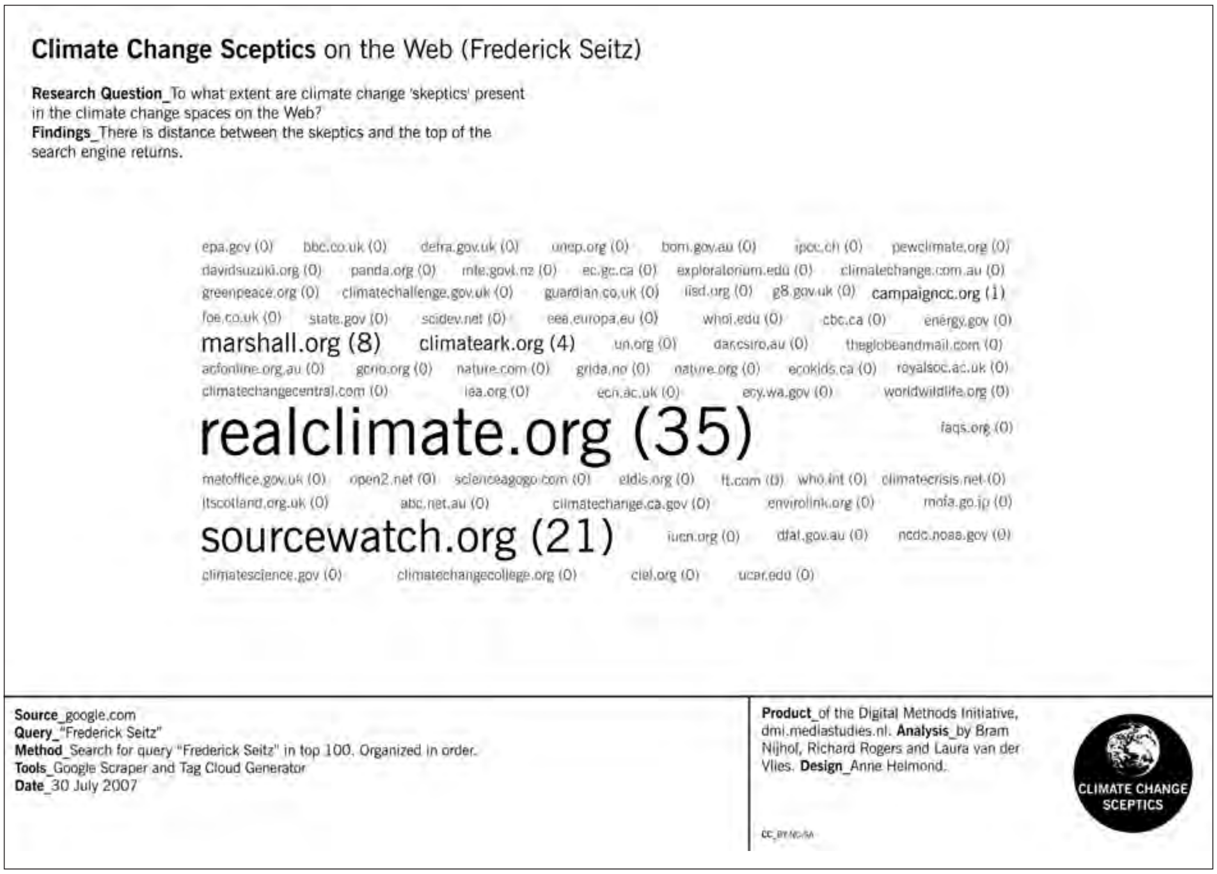


Figure Four: Extended follow-follower network of Dutch Ministry-level civil servants, March, 2013. Data captured by TCAT, DMI Amsterdam, and Visualization by Gephi. Source: Baetens et al., 2013.



Climate Change Sceptics on the Web (Frederick Seitz)

Research Question_To what extent are climate change 'skeptics' present in the climate change spaces on the Web?
Findings_There is distance between the skeptics and the top of the search engine returns.

epa.gov (0) bbc.co.uk (0) defra.gov.uk (0) unep.org (0) bom.gov.au (0) ipcc.ch (0) pewclimate.org (0)
davidsuzuki.org (0) panda.org (0) mfe.govt.nz (0) ec.gc.ca (0) exploratorium.edu (0) climatechange.com.au (0)
greenpeace.org (0) climatechallenge.gov.uk (0) guardian.co.uk (0) iisd.org (0) g8.gov.uk (0) campaigncc.org (1)
foe.co.uk (0) state.gov (0) scidev.net (0) eea.europa.eu (0) whoi.edu (0) cbc.ca (0) energy.gov (0)
marshall.org (8) climateark.org (4) un.org (0) dar.csiro.au (0) theglobeandmail.com (0)
acfonline.org.au (0) gcrio.org (0) nature.com (0) grida.no (0) nature.org (0) ecokids.ca (0) royalsoc.ac.uk (0)
climatechangecentral.com (0) iea.org (0) ecn.ac.uk (0) ecy.wa.gov (0) worldwildlife.org (0)
realclimate.org (35) faqs.org (0)
metoffice.gov.uk (0) open2.net (0) scienceagogo.com (0) eldis.org (0) ft.com (0) who.int (0) climatecrisis.net (0)
ltscotland.org.uk (0) abc.net.au (0) climatechange.ca.gov (0) envirolink.org (0) mofa.go.jp (0)
sourcewatch.org (21) iucn.org (0) dfat.gov.au (0) ncdc.noaa.gov (0)
climatescience.gov (0) climatechangecollege.org (0) ciel.org (0) ucar.edu (0)

Source_google.com
Query_"Frederick Seitz"
Method_Search for query "Frederick Seitz" in top 100. Organized in order.
Tools_Google Scraper and Tag Cloud Generator
Date_30 July 2007

Product_of the Digital Methods Initiative, dmi.mediastudies.nl. **Analysis**_by Bram Nijhof, Richard Rogers and Laura van der Vlies. **Design**_Anne Helmond.

CC_BY-NC-SA

CLIMATE CHANGE SCEPTICS

Figure Five: Climate change skeptics' presence in the leading climate change websites, according to google.com, July 2007. Source distance analysis by the Google Scraper, aka the Lippmannian Device. (cc) Digital Methods Initiative, Amsterdam, 2007.

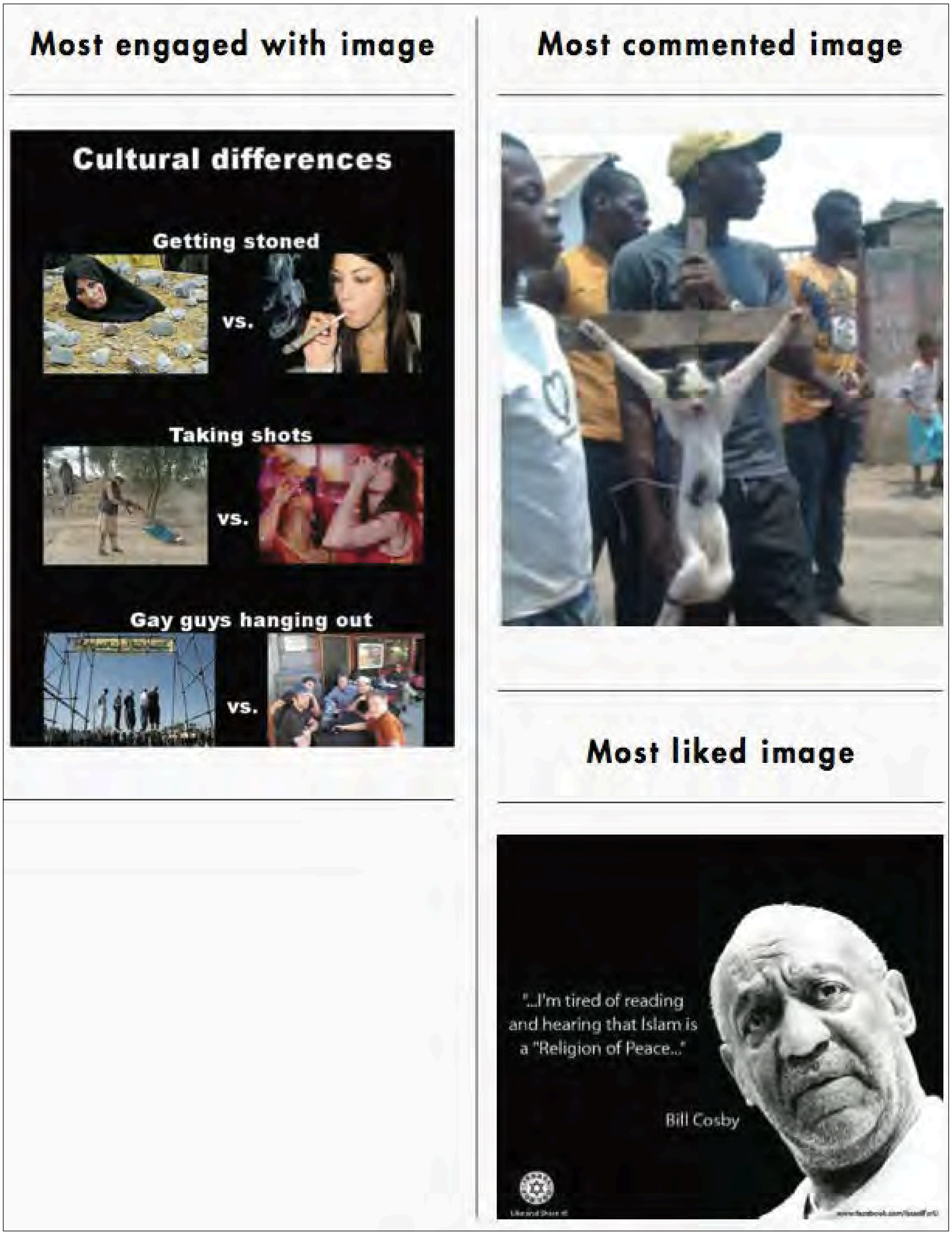


Figure Six: Most engaged with content in European counter-jihadist networks on Facebook, January 2013. Product of "What does the Internet add? Studying extremism and counter-jihadism online," International Workshop and Data Sprint, (cc) Digital Methods Initiative, Amsterdam, 2013.

NECSUS European Journal of Media Studies

Spring 2014_'Traces'

Published:

June 25, 2014

Tags:

consumer surveillance, cookies, digital methods, traces, Web tracking

The Third Party Diary: Tracking the trackers on Dutch governmental websites

by Lonneke van der Velden

Introduction: Web tracking as data and Web tracking as issue

There are a range of 'privacy enhancing' tools on the Web. In this article I will discuss how the browser plugin Ghostery transcends individual usage. By making Web tracking transparent it empirically and conceptually contributes to a particular understanding of contemporary consumer surveillance.

Ghostery detects techniques (called 'third party elements') that collect data on Internet users when they visit certain websites; Ghostery also gives the user an alert with a small visualisation in the Web page. The fact that Ghostery has specific detection principles makes the tool useful for Web researchers as well. Building upon the work of the Digital Methods Initiative (DMI) which specialises in repurposing Web devices for research I have explored the 'Tracker Tracker'.[1] The Tracker Tracker mobilises Ghostery's capacities for the study of third party elements on specific sets of URLs. In this way it enables the comparison of the presence of third party elements in a more systematic manner.

In my case study I used this tool to look more closely into a sample of Dutch governmental websites in 2012. The reason for doing this case study was twofold. First of all, online tracking by Dutch governmental websites was controversial at the time. There was discussion about the Dutch implementation of the EU e-Privacy Directive and the extent to which the Dutch government was still tracking Internet users without their consent, hence failing to obey the law. My question was whether it was possible to measure the governmental response to this debate by using the Tracker Tracker to map the presence of third parties on governmental websites over time. The results pointed to an average of almost 60% presence of third parties and indicated that the government responded only slowly, if at all, to the affair. The results also showed clusters of websites sharing similar third parties. This raises questions about the way governmental websites perform different roles online; in addition to their expected and visible role as the main public service providers they also have an active role in contributing to the information economy by sharing (personal) data with major corporations.

A second reason for using the tool in the context of a particular Dutch local affair was that it was a way of 'situating' Digital Methods. This should be seen as a more experimental attempt to discuss how the Tracker Tracker tool performs in relation to a particular data set. Some of my results made me think about Ghostery's method of working and its capabilities, an issue that links up to wider academic debates about the increasing role of digital devices in social research.[2] The Digital Methods program mobilises digital devices explicitly for knowledge production. However, as Marres & Weltevrede argue, devices come with 'epistemology built in'.[3] This subsequently also raises questions about the politics of knowledge that these devices bring along, questions that a variety of digital methods researchers are currently examining.[4] For example, Marres has questioned the kind of methods that are remediated by Web devices and how that affects the work that comes out of the research assemblage in which these devices participate.[5]

Ghostery also lends itself to a more in-depth inquiry; it is an example of a device that brings Web tracking into view in order to make Internet users aware of the fact that their browsing behaviour is being monitored. That means that Ghostery is implicated in a particular issue and uses a specific repertoire to explain what Web tracking is about. Therefore an important question arises about what way Ghostery brings this issue to the fore.

Digital devices in action

This article has a running concern with how Ghostery brings Web tracking into view and what that means for the way it participates in the research project. According to Gitelman & Jackson, data are often imagined as being picked up from some 'undifferentiated blur'. In many discourses data are talked of as being 'collected', 'piled', or 'mined'. However, as these authors go on to argue, data always depend on operations of knowledge production. Data, as they quote Lev Manovich, do not just 'exist' but need to be 'generated'.[6] In *Raw Data is an Oxymoron*, Gitelman & Jackson aim to pursue the question of 'how different disciplines have imagined their objects and how different data sets harbor interpretative structures of their own imagining'.[7]

When using Web devices for research a reframing of this concern would be a need to consider how these devices imagine data and how this feeds back into our data sets. The specific use of the term 'device' by Ruppert & Law & Savage is useful here. They state: '[w]ithin these cascades [of applications and software] a device can make, compile and transmit digital data and/or remake, analyse and translate data into information and interventions.'[8] They stress the organisational activity of devices in which both knowledge and social action get distributed. By doing so devices are constitutive of emergent social relations. Similar to the performativity of devices of the social sciences and economics,[9] say Ruppert et al., digital devices 'enact' the social. They 'inscribe' something into the very thing they attempt to analyse. This is a reason for them to say that key to what we as digital researchers ought to do with regard to digital devices is to get close. That is, to

> get our hands dirty and explore their affordances: how it is that they collect, store and transmit numerical, textual, aural or visual signals; how they work with respect to standard social science techniques such as sampling and comprehensiveness; and how they relate to social and political institutions.[10]

As I hope to illustrate, Ghostery proves to be a good opportunity for such an exploration. I will look at the context in which it operates, its method, assumptions, affiliations, and suggestions for actions, and how that is constitutive for the issue of online tracking. In line with other work in science and technology studies (STS) I will look at the 'situated, material conditions of knowledge production'.[11] In other words I will first approach the device as an 'object' of study before repurposing it as a 'method' for research, a distinction made in the work by Marres & Weltevrede.[12] Another way of putting it would be that this is an investigation into a 'device in action'. By setting the study up in this way there will be several instances in which the generation of data is made explicit. I discuss how Ghostery imagines data, how the output of the Tracker Tracker tool shows in what ways third parties get their data, and how I treated the data set myself. In all these moments I try to show how data is organised through different formats and how these formats, in the context of the case study, interact.

Getting close to Ghostery

Ghostery operates in the context of a data market in which website optimisation coincides with behavioural advertising. Webmasters make use of corporate tools to keep track of their visitors and often share the data with third parties, for example advertising networks. As McStay explains: '[b]ehavioral advertising tracks users' browsing activities between websites over a period of time for the purposes of serving advertising tailored to what advertisers assume are users' interests.'[13] These assumed interests are extracted from the type of websites and other indicators of browsing behaviour (such as location, time, type of device, etc.). After the data are collected, stored, and aggregated, profiles are sold at real-time biddings. Advertisers can bid for advertising space delivered to specific users – the more detailed the profile the higher its value.[14] Just as in the 'regular' financial sector this market comes complete with 'data brokers' and 'data speculation'.[15] To characterise the culture of data trade metaphors such as 'Data Wild West' circulate among marketers themselves as well as among their critics.[16] For individual users it is not easy to know what happens with data that are collected because the privacy policies of companies are not very transparent.[17]

In this context a range of tools are developed that tell users that their online behaviour data is being monitored.[18] To give a few examples: Lightbeam (previously called 'Collusion') is a Firefox browser plugin developed by Mozilla that will display your online traces through a real-time network graph; another tool is Disconnect, a Chrome extension that will visualise third party trackers per site you visit and provide you with a bar chart estimating the time that you saved yourself if you decided to block the trackers. Ghostery, which is the central actor in this study, delves deep into the trackers. Whereas privacy policies that are supposed to clear up what happens to user data remain opaque Ghostery brings the instruments that are crucial in this process into view. As stated on the website, it 'shows you the invisible web – cookies, tags, web bugs, pixels and beacons – and gives you a roll-call of over 1,800 ad networks, behavioral data providers, web publishers and other companies interested in your activity'.[19]

Ghostery is above all a visualisation tool that focuses on the *collectors* of data; it makes a translation from pieces of code in the page source to the specific type of tool it recognises this code to be a trace of. For example, 'http://b.scorecardresearch.com/beacon.js?_=1391171393485' is recognised as 'ScoreCardResearch Beacon'. Ghostery proceeds to bring this finding to the screen by displaying a pop-up. In the screenshot below you can see that when one visits this particular website (of the police) there are also two third parties present: Google Analytics and ShareThis. In this particular example Ghostery shows that this computer is not only communicating to the server of the website but also to the servers of other third party companies.

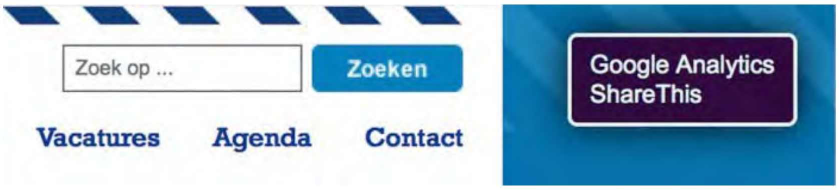


Fig. 1: Pop-up Third Parties, http://kombijdepolitie.nl, January 2014.

To describe the techniques that collect user data Ghostery uses the term 'third party elements', or in short '3pes'. Ghostery orders and ranks third party elements by indexing them into different types. It does so not according to their technological terms (such as pixels and bugs) but according to what they *do*. Ghostery says third party elements can deliver advertisements (AD), provide research or analytics for website publishers (AN), track user behaviour (T),[20] provide some kind of page function through widgets (W), or disclose data practices involved in delivering an ad (P).

Ghostery's ranking system (Ghostrank) presents the weight of these elements according to their relative presence on the Web – at least on the part of the Web that is visited by Ghostery's user population, because Ghostrank is made possible through the participation of the people who use the tool. The database is constructed by people that opt-in to automatically share their third party encounters with Ghostery's database. In spring 2013 Ghostery had 17 million users and 7 million took part in Ghostery's 'panel' that contributes to the database.[21] The table has the form of the periodic table of elements (http://knowyourelements.com). The higher the relative chance one encounters a specific third party element the higher it is ranked in the table. Therefore by providing visualisations and information during browsing Ghostery makes third party elements not only 'present' but also more accessible for further analysis.

By making the invisible Web visible Ghostery aims to help Internet users to make informed decisions and to give them more control over when they are being tracked and by whom. The behaviours per element are filed in a library. According to Ghostery's parent company, Evidon, the library contains more than '1,600 companies and 4,100 different types of trackers', which makes it, according to them, 'the only comprehensive library of trackers on the internet'.[22] The library provides information about what kind of data are collected (such as geo-location data, IP address, or phone number) by a particular third party element and whether it shares data with (again) other parties. Ghostery also suggests different ways to 'handle' third parties. It offers users the possibility to block all or only some third parties by separately flagging them.

The database is not only of use to privacy-aware individuals. Evidon uses the information to inform online marketing companies about the implementation of their tools and to offer advice about how to comply with privacy rules.[23] Evidon's mission is 'to enable a more transparent, trusted environment for consumers and advertisers'.[24] The company takes part in a larger program managed by a consortium of advertising and marketing associations – the Digital Advertising Alliance (DAA) – which pushes a label that draws a parallel with ethical (food) consumption, referring to the idea of a nutrition label: '[f]or businesses and NGOs, Evidon provides the technological underpinnings that put the AdChoices icon, which functions as a "tracking nutrition label" into ads, as well as reports on trackers and what they are doing on the web.'[25]

Ghostery as an issue device

Now that we have gotten to know Ghostery a bit better we can get back to how to think about 'devices in action'. How does Ghostery (following Ruppert et al.) distribute information and intervention, and what does that inscribe to the issue at hand? Through its database and vocabulary Ghostery mobilises particular concepts and distributes what counts as information and action. Through Ghostery Web tracking becomes something that can be ordered, something that becomes knowledgeable. The (visual) language of the periodic table is maybe just a metaphor but at the same time it helps framing trackers as components and tracking as an environment. Trackers, instead of consisting of intangible processes, become elements that can be mined themselves.

From science and technology studies we know that ideas of nature can be constitutive in sorting out what belongs to the realm of knowledge and what belongs to the realm of values (and social action).[26] Ghostery is engaged in a similar distribution as well – in addition to the third party environment as something to become 'informed about' one can also learn how to 'cope' with it. By offering a knowledge repository accompanied by an action repertoire of possible 'options' you can detect, block, and pause. There is a common denominator in this action repertoire – 'you'. How to evaluate Web tracking becomes a matter of responsibility on the part of the individual Internet user, who can asses his or her trust relation with different kinds of companies. Tracking becomes something that the info-aware individual can choose to consume or not.

In a text on data communities Harvey et al. use the notion of 'transparency devices' to describe how these communities map things such as government transactions or community conflicts with a set of specific tools for measurement and visualisation; but they also show how these communities, by making things transparent and legible, simultaneously inscribe something to the thing they study.[27] Ghostery does exactly that. Through making Web tracking transparent it enacts tracking as a material thing, as something consisting of components that can be studied and ranked; it subsequently calls an ethics of Web tracking into existence. Web tracking can be 'bettered' through labels, changing consumer behaviour and coalitions between companies. Thus, in addition to looking at community practices we can also analyse processes by which transparency and inscription coincide through devices themselves. Here I refer to the work of Marres who has coined the term 'material participation'.[28] With this term she wants to stress the extent to which objects can facilitate matters of concern, and 'issue articulation' is one way in which this happens. Building upon Marres' work we could say that Ghostery is a device that 'redistributes participation'; by articulating the issue in this way it organises the work and responsibilities relating to how to cope with Web tracking. So, if digital devices materialise social relations, Ghostery materialises an issue and it does so in a very literal sense. Therefore I use the term issue device rather than transparency device to refer to the way in which Ghostery brings Web tracking to the fore, because I think the performativity of the issue is a relevant point if one is concerned with what the device does to the method.

Ghostery as a research device

Because Ghostery provides certain ordering principles to detect third parties and a typology relating to their activities it has proved to be very useful as a research tool. The Digital Methods Initiative at the University of Amsterdam deploys the ordering principles of existing Web devices for social research. Considering that these devices take part in specific 'device cultures' they can produce situated knowledge that is valuable for understanding contemporary social life.[29] The adagio is to follow the 'language of the medium', or the 'actors', in Latourian jargon.[30] That means instead of using previously established categories from the social sciences that emerged out of other research sites besides the Web, one would stay close to the terms of Web devices and look at how they articulate the connections between various Web objects.

The Tracker Tracker is part of the toolbox of the Digital Methods Initiative. The Tracker Tracker uses a database of pre-defined fingerprints of Web technologies provided by Ghostery and compares those traces with the URLs that are of interest to the researcher. The DMI built upon Ghostery and not on a comparable device such as Lightbeam because the latter was not yet publicly known at the time the Tracker Tracker tool was built, also because Ghostery publishes their lists of trackers and updates them regularly. This enables researchers to analyse specific data sets by making use of Ghostery's classificatory scheme. After inserting a list of URLs into the Tracker Tracker it provides a spreadsheet with all the domain names and the respective names of third party elements that are detected per URL, also adding their type (AD, Analytics, Widget, etc.). Therefore the tool does not only give an indication of the overall presence of third party elements that collect data online but it also enables you to zoom in on the different types of elements and to do a comparative analysis between websites.

Tracker Tracker research has been relatively new and experimental; projects have been done with data sets such as the top-Alexa websites, technology blogs, and political party websites.[31] As work by Gerlitz & Helmond on the top-1000 Alexa websites has shown, the Tracker Tracker can be used to map the connections between websites and the 'data objects' that they share. Such maps provide insight into what they call an 'alternative fabric of the web'. This texture is not based on the hyperlinks through which we often imagine the Web but on the relations between third party tracking devices and the respective websites at which they are detected.[32] If we look at such networks of websites we get a glimpse of the material relations that provide the conditions for data transactions within the previously mentioned 'Wild West'. Hence, this kind of exploratory research helps us to imagine the contribution of data collectors to what Callon & Muniesa have termed 'calculative spaces' – those arrangements that make things calculable.[33] In line with these kinds of digital methods studies, I looked at the shared third party elements on a particular set of websites, particularly the websites of the Government of the Netherlands.

The Third Party Diary

The context of my case study was a debate in The Netherlands about the Dutch implementation of the EU e-Privacy Directive. Since June 2012 the Dutch law obliges website owners to ask for the consent of Internet users for technologies that access their devices in order to collect or store data – a law which became (badly) known as the 'cookie-law'.[34] A few months later the Government of the Netherlands ('Rijksoverheid') was criticised for failing to obey the law. The debate focused on two main governmental websites: rijksoverheid.nl and government.nl. Both sites were setting cookies. On 9 August 2012 the government announced that they would disable all the cookies on these two websites and that they would further assess whether 'other websites' needed to be adjusted as well.[35]

This discussion provided an incentive for me to dig a bit deeper into this issue. The response by the government made me think about which 'other websites' could be of relevance. Thanks to open data guidelines the whole Website Register of the Government of the Netherlands ('Websiteregister Rijksoverheid') can be found online. This register gives information about approximately 1100 websites that belong to the Dutch government (cities and regional governments are excluded).[36] This data set provided the starting point for my research. The question about which particular tracking devices are allowed (or not) I will leave aside by reformulating the debate in socio-technical terms: can we measure the response of the Dutch government to this issue to by mapping the presence of third parties on these websites?

For four months in 2012 I registered the third parties that collect visitors' data on websites belonging to the Government of the Netherlands. I presented the results in an online logbook titled The Third Party Diary, which gives an impression of third party encounters when visiting the government online (http://thirdpartydiary.net). The format of the diary was chosen for several reasons. Keeping a diary would be a means to structure the project and feature the results online, as it dealt with a current affair.[37] Another reason was that the research was not a clean and automated process and I did not want to suggest it was – working with this device was in fact pretty messy.[38] As argued by Leistert, digital methods can give the impression of being some kind of disembodied process with respect to the objects of research and the researcher as well.[39] A diary seemed to be a good format to deal with the idea that the outcome of the project was not just through the tool but also through an engagement with the tool.

The methodological steps I took were as follows. I inserted the total list of URLs in the Website Register in the Tracker Tracker tool. The Tracker Tracker output mentions third parties multiple times per domain name when similar elements are detected in different 'patterns'. Therefore these double findings were deleted from the tool's results. I then determined the total list of domain names containing third party elements, the total amount of third party elements, and I randomly checked for false positives and negatives. I repeated the study every month for four months, from August until November 2012. In 2013 the study was taken up again in January and repeated irregularly. The Website Register of the Government of the Netherlands is regularly updated. The latest revision of the register was used as input for the Tracker Tracker tool each time. Below I will present my findings and discuss how this contributes to an understanding of Web tracking practices.

Third party presence

In August 2012, in total, 856 third party elements were detected by 38 different individual third parties (Google Analytics, Webtrends, Facebook Connect, etc.). The figure below is a visualisation of the relative presence of third party elements (the size refers to the amount of third party elements, the colour to the type of activity).

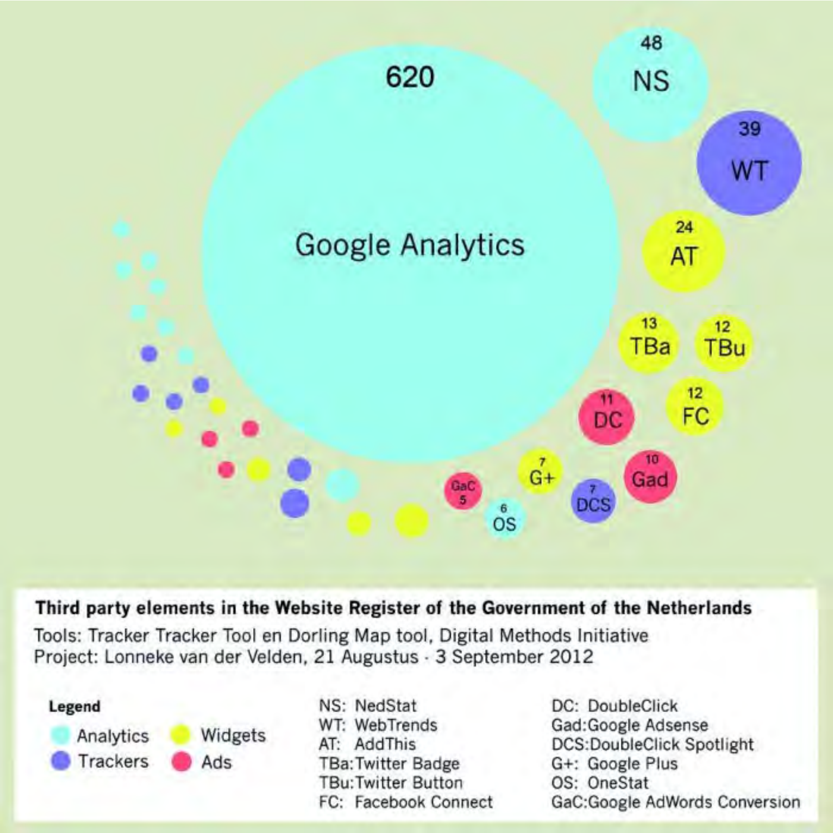


Fig. 2: Third party presence, August 2012. The nodes refer to the different third party elements (3pes) as distinguished by Ghostery (http://www.knowyourelements.com/). Elements that occurred less than five times are not listed by name. The size indicates the amount of 3pes in the Website Register of the Government of the Netherlands and the colour refers to the type of 3pe. The Register contained 1110 websites in total.

Several third party elements are operated by the same company, which leads to the conclusion that only 28 companies seem to be involved, of which Google is the biggest (see Figure 3 below) followed by Comscore, Webtrends, Twitter, AddThis, and Facebook. This finding is supported by Hoofnagle et al., who reviewed tracking practices on top websites in 2009 and 2011 and concluded that there is a concentration of a relatively small amount of companies operating a large amount of Web tracking technologies.[40]

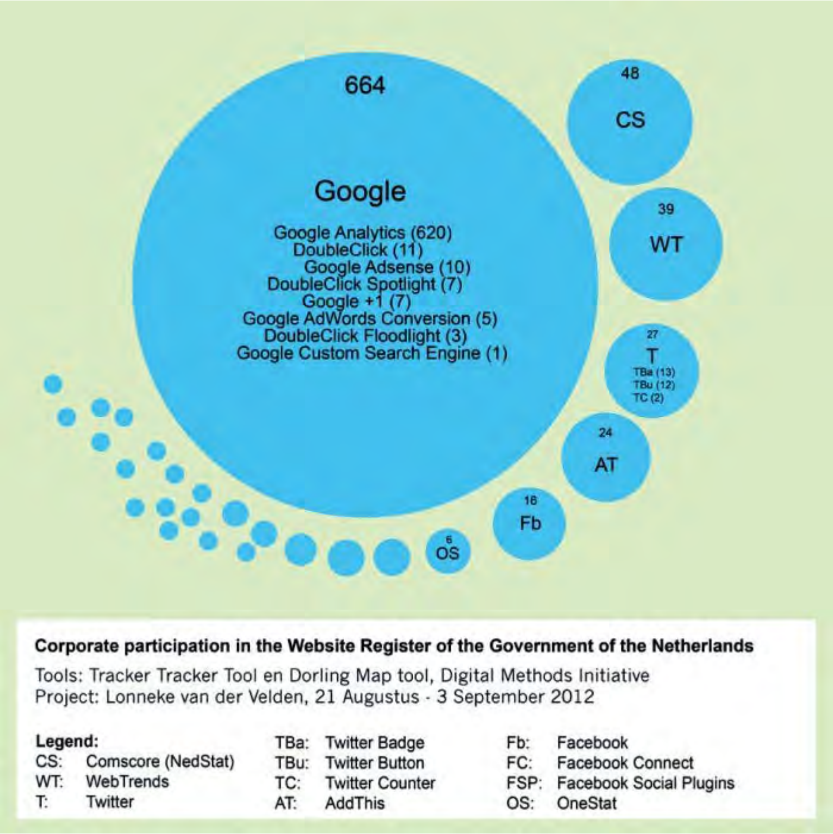
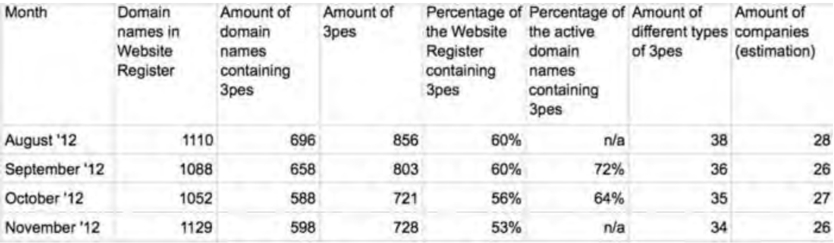


Fig. 3: Corporate participation, August 2012. The nodes refer to third party elements (3pes) in the Website Register of the Government of the Netherlands, as indicated by Ghostery (http://www.knowyourelements.com). Elements that occurred less than five times are not listed by name. The size indicates the share in 3pes companies have in the total amount of 856 3pes. The register contained 1110 websites in total.

On average the percentage of websites containing third party elements is always more than half of the website register. The percentage lies higher when taking into account the fact that many domain names are not even active. For instance, in September the Website Register contained 1088 websites of which 913 were active. 658 domain names contained third party elements – that makes 60% of the whole register but 72% of the active domain names. A study by Koot, who simultaneuously investigated the same data set as I did in September 2012 (though using a different approach), points to similar findings. He used software for automated browsing (Mozrepl and Burp Suite) in order to fetch the third party content on the domain names and to analyse the traffic.[41] He found that 671 domain names of the active URLs contained third party content (73%). Thus, despite Ghostery's detection method not being 100% complete[42] it does come pretty close to the findings of other researchers.

Table 1 gives an overview of the presence of third party elements in the website register for the months August-November 2012, the months directly following the public debate.



Month	Domain names in Website Register	Amount of domain names containing 3pes	Amount of 3pes	Percentage of the Website Register containing 3pes	Percentage of the active domain names containing 3pes	Amount of different types of 3pes	Amount of companies (estimation)
August '12	1110	696	856	60%	n/a	38	28
September '12	1088	658	803	60%	72%	36	26
October '12	1052	588	721	56%	64%	35	27
November '12	1129	598	728	53%	n/a	34	26

Table 1: Results 3pes (August-November 2012).

Because the government was given an explicit warning in September 2012 by the Independent Post and Telecommunications Authority of the Netherlands (OPTA) to abide by the law, I expected to see a decrease in third parties over time.[43] There was a small drop in October and November but it is hard to say whether that really indicates removal. The decrease might also be due to the fact that the Website Register was updated and now excludes a few redirects that were included in September.[44] In November 2012 the overall percentage of third party elements in the Website Register was still 53%. Hence, over four months the decrease in third party elements was 7%. In fact, when I checked a year later in December 2013 the percentage was back to 63%. We can therefore conclude that after the August 2012 debate about the government tracking their visitors the removal of tracking devices has been limited.

Shared third parties

It is also possible to visualise the connections between websites and third parties. The image below gives an impression of the associations between the third party elements (the collectors of the data) and the websites within which the elements are located. The output of the Tracker Tracker tool from September 2012 was visualised with Gephi.[45] It shows the massive outreach of Google Analytics; it also shows how certain nodes are surrounded by clusters of websites, for instance the Webtrends cluster on the bottom right. This means that several websites use a Webtrends tracker.

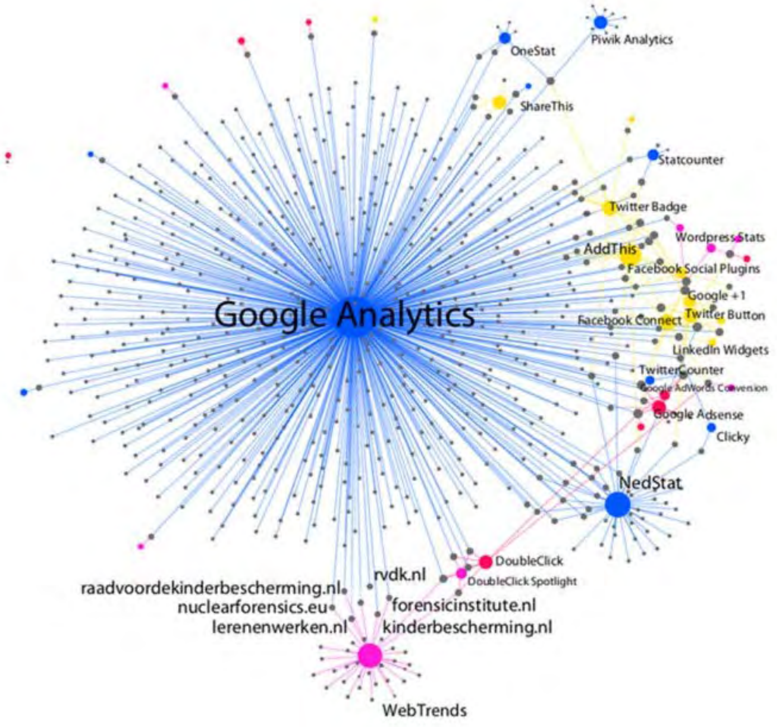
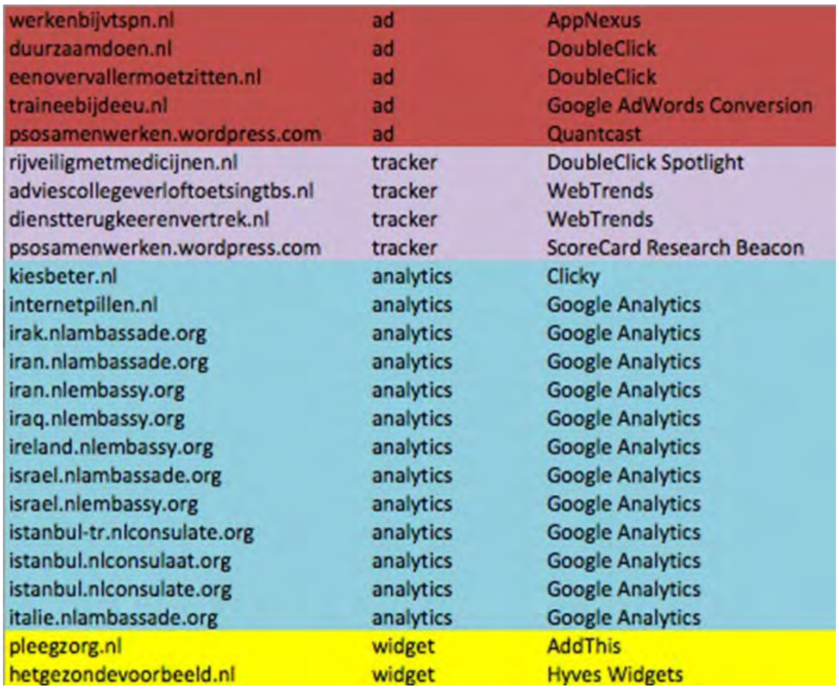


Fig. 4: Gephi visualisation, September 2012. The coloured nodes are trackers. The grey nodes are the domain names. The names of the websites are deleted for reasons of clarity, except for the bottom to illustrate the purpose of the map. For instance, nuclearforensics.eu and forensicinstitute.nl are connected with both WebTrends and Google Analytics.

There are a few interesting insights when zooming in further into that particular cluster. I first manually sorted the results by 3pe-type and name (see Table 2 below).



werkenbijvtspn.nl	ad	AppNexus
duurzaamdoen.nl	ad	DoubleClick
eenovervallermoetzitten.nl	ad	DoubleClick
traineebijdeeu.nl	ad	Google AdWords Conversion
psosamenwerken.wordpress.com	ad	Quantcast
rijveiligmetmedicijnen.nl	tracker	DoubleClick Spotlight
adviescollegeverloftoetsingtbs.nl	tracker	WebTrends
dienstterugkeerenvertrek.nl	tracker	WebTrends
psosamenwerken.wordpress.com	tracker	ScoreCard Research Beacon
kiesbeter.nl	analytics	Clicky
internetpillen.nl	analytics	Google Analytics
irak.nlambassade.org	analytics	Google Analytics
iran.nlambassade.org	analytics	Google Analytics
iran.nlembassy.org	analytics	Google Analytics
iraq.nlembassy.org	analytics	Google Analytics
ireland.nlembassy.org	analytics	Google Analytics
israel.nlambassade.org	analytics	Google Analytics
israel.nlembassy.org	analytics	Google Analytics
istanbul-tr.nlconsulate.org	analytics	Google Analytics
istanbul.nlconsulaat.org	analytics	Google Analytics
istanbul.nlconsulate.org	analytics	Google Analytics
italie.nlambassade.org	analytics	Google Analytics
pleegzorg.nl	widget	AddThis
hetgezondevoorbeeld.nl	widget	Hyves Widgets

Table 2: Third party elements sorted by type, September 2012 (Selection. Complete list available at http://thirdpartydiary.net).

It is here that Ghostery becomes more than a magnifier and shows its microscopic capacities. This way of sorting shows which websites share similar third party elements and how in some cases the use of third party elements corresponds to departmental orderings of the respective ministries. For Table 2 I selected only a sample, but at least 23 sites of the Ministry of Security and Justice were using Webtrends in September 2012, including sites such as the website of the Council for Child Protection (Raad voor de Kinderbescherming), a committee for research into child abuse (Commissie-Samson), and a committee advising on the release of mentally-disordered offenders (Adviescollege Verloftoetsing TBS).

Trying to zoom in even further I picked one website, the website of the Council for Child Protection (kinderbescherming.nl), and received a Webtrends cookie in my browser which included my IP address. The IP address stayed the same when I visited another website of the Ministry of Security and Justice (avtminjus.nl) within the Webtrends cluster. Webtrends only set a new cookie when I emptied my browser. Checking the host of these Webtrends cookies led me to a company called Imetrix, which provides hosting and analytics. Apparently the Ministry of Security and Justice hired this company to take care of a whole set of its websites.[46] This suggests Imetrix collected IP addresses (and maybe more data) categorised in a specific 'departmental' way, through websites that deal with child protection issues and mentally-disordered offenders – issues that fall under the category of 'Security and Justice'. They removed the trackers by the end of 2012.

Another interesting insight from the same data set is that all Dutch embassy websites share Google Analytics. In Ghostery's library one can find a summary of what Google Analytics collects, which includes (according to their terms) anonymised IP addresses, locations, and search queries. This means that this kind of information related to people interested in Dutch embassies is most probably shared with Google's servers. The cluster entails 250 Dutch embassies and consulates. The point here is not only that behavioural data is transferred from governmental websites to third parties, but it is the standardisation in this process that raises interesting questions. Because the government implemented Google Analytics as a *standard* on almost all of the ambassadorial websites the government shared with Google a data set that is in effect organised (as an ambassadorial category), and as the December 2013 results indicate they still did so a year later.

Lessons from The Third Party Diary

The results of the case study raise critical political-economic, legal, and security-related questions. Is the Dutch government, in a sense, a 'miner' for what Leister calls 'Wild West data mining capitalism',[47] by already preparing datasets and giving companies such as Google and Facebook a helping hand in 'audience sorting'? And since we are already familiar with Google Flu Trends as a form of research into flu activity (http://www.google.org/flutrends/) one could imagine what kind of 'trends research' Google could do with ambassadorial data sets. Will 'Visa Request Trends' become the new migration studies? There could be potential legal consequences as well, because data is shared with servers that are under the jurisdiction of the United States. More concretely, the use of tracking devices can bring along a range of privacy and security problems. Koot's study explains how third party content can provide easy access points for cyber attacks (such as session hijacking and malware infection).[48] Tracking devices can be 'repurposed' too. Since the leaking of the NSA files by Edward Snowden we know that Google cookies are repurposed by the NSA to follow the behaviour of potential targets before the agency installs malware on their computers.[49] These new insights into the use of Web tracking devices show how consumer surveillance and state surveillance coincide.

The case study raises questions with respect to the method as well. Over time a few websites changed their tracking policy and began to ask for explicit consent from the visitor (for example the Education Council of the Netherlands at http://onderwijsraad.nl). This basically means that the Internet user will get a pop-up that asks whether he or she agrees with the use of cookies. Upon agreement the page should load the trackers, or otherwise it should not (ideally speaking). The effect of this change was that some third party elements disappeared from my output. However, this does not mean that third party elements are not operative. Studies have shown that people tend to accept terms of services.[50] Therefore people may consent to and load third parties that were (at the time of the project) not indexed by the Tracker Tracker output of Dutch websites. The disappearance of third party elements is therefore an interesting phenomenon by itself.

Elmer, about a decade ago, argued that cookies should be understood as mechanisms of communication instead of using the flattened definition of 'a piece of text'. According to him the 'data definition' of cookies obscures the process by which this information reaches the hard drive of the computer.[51] Along the same lines, in the example of the webpage above, the loading of third party trackers also depends on a process of negotiation. Moreover, the way websites organise the consent-procedure happens through different programming languages. At the time of study the Tracker Tracker tool did not recognise JavaScript and therefore behaved as an atypical and old-fashioned browser. Some websites will treat this as a 'yes I accept' and others as a 'no'. In other words, the device cannot consent. It is treated differently depending on how the website treats the device.

This brings me to the more reflective question of whether turning Ghostery from issue device into research device mattered for the way Web tracking was presented in the research project. What is, to recall Marres & Weltevrede, the epistemology built into the tool? Does it matter that Ghostery imagines 'tracker data' as components, as a materialised environment, as things that can be mined in turn, and that it distributes 'tracker allowance' to the realm of individual choice? To a certain extent I think it does. If we follow the device by only focusing on its detection principles we limit ourselves to an elementary understanding of tracking in which it is located in the page source. The Tracker Tracker then operates under the assumption that the activities of third party elements are dictated by the set of sites and their code. However, we cannot assume that in this case.

Since we are dealing with a particular local context in which website owners are encouraged to ask for consent and people have to interact with that code the social or legal-material arrangement is one in which interventions take place before scripts are loaded. In some of these cases, depending on how the website responds to the Tracker Tracker's automated character and the inability of the tool to interact with site content like a regular visitor, it will not show all the trackers the latter would encounter. A negative output from the Tracker Tracker tool cannot be judged 'tracker clean' unless a manual check – by accepting cookies – follows. In other words, in this context 'tracker allowance' turns out to be more complex than individual choice only because Web tracking is dealt with through a complex of state legislation, cookie-walling, and user interaction. This becomes particularly relevant in research projects with smaller and specific data sets. A question of methodological challenge then becomes whether it is feasible for digital methods to enrich the Tracker Tracker in such a way that it captures these processes of negotiation and acceptance. Can 'docility' be built in?[52] At the time of writing an update of the tool is being worked on (in the sense that it now recognises JavaScript).

Lury & Wakeford have compiled a range of studies on devices clustered under the term 'inventive methods'. According to them devices can be inventive when they can 'change the problem to which they are addressed'.[53] In this case study the Tracker Tracker has prompted a reorganisation of the project by provoking new questions: can we capture Web tracking as a more interactive thing? Should and can the tool be changed in order to do that? A more general conclusion for future tracker research could be that the context of the data set matters. One could use digital methods to study 'social life' (in my case this was the state of the issue and institutional-tracking assemblages). However, it is important to ask what kind of new questions a data set brings with respect to the Web objects that we investigate.

Conclusion

In this project Ghostery was shown to operate on a range of levels. As an issue device it brings Web tracking to the fore, and we need a qualitative approach to see that. Ghostery maps and ranks practices of Web tracking and uses a particular vocabulary to make these technologies present and accountable. Ghostery's inscription into the issue is one in which Web tracking becomes a material environment to be coped with.

As a research device it can point out the associations between websites and shared objects and relate to existing studies into the transactions of behavioural data. The Tracker Tracker also allows zooming into clusters of websites and provides empirical data that can feed concrete public affairs. The Government of the Netherlands was shown to intensively participate in the market of behavioural data. We get some insight into how specific data move from one organisation to another, such as from the ministry of Foreign Affairs to Google. It gives few clues about the make-up of these data sets and about which actors participate in this process. The project therefore contributes to a better understanding of the first steps of the process of behavioural targeting. It suggests that orderings by category are already embedded in the process of collecting data due to very mundane and institutional aspects of governmental life. Thus, instead of assuming that data collection is a starting point for further enhancement and profiling processes, practices of categorisation turn out to be already active from the start.

The case study has also interrogated the device. Reflecting upon the way Ghostery imagines its data and taking the device out of its device culture to study a new context has led to the question of how to capture Web tracking as a negotiated practice.

Author

Lonneke van der Velden is a doctoral researcher in the Digital Methods Initiative (DMI) at the University of Amsterdam. She is interested in issues of surveillance, publics, and evidentiary technologies. More particularly, she looks at initiatives that turn surveillance into an object of public scrutiny through the use of digital tools. Van der Velden has a background in Science and Technology Studies and Philosophy and is part of the editorial board of *Krisis*, a Dutch open access peer-reviewed journal for contemporary philosophy.

Acknowledgements

I would like to thank the anonymous reviewers for their comments as well as the participants at the workshops organised by the Digital Methods Initiative (DMI) and Goldsmiths College, particularly Noortje Marres. Matthijs Koot helped me with his topical expertise and my colleagues at the DMI, in particular Erik Borra and Emile den Tex, supported me in better understanding the technicalities of the tool.

References

Berg, B. van den and Hof, S. van der. 'What happens to my data? A novel approach to informing users of data', *First Monday*, Vol. 17, No. 7, 27 June 2012.

Borra, E. K. and Rieder, B. 'Programmed Method. Developing a Toolset for Capturing and Analyzing Tweets' in *Aslib proceedings*, edited by K. Weller and A. Bruns, 2014.

Brock, J. 'Credibility Gap: What does Ghostery really see?', *Privacy Choice Blog*, 4 March 2010: http://blog.privacychoice.org/2010/03/04/credibility-gap-what-does-ghostery-really-see/ (accessed on 25 January 2014).

Callon, M. and Muniesa, F. 'Peripheral Vision: Economic Markets as Calculative Collective Devices', *Organization Studies*, Vol. 26, No. 8, 2005: 1229-1250.

Elmer, G. *Profiling machines: Mapping the personal information economy*. Cambridge: MIT Press, 2004.

Gerlitz, C. and Helmond, A. 'The Like Economy: Social Buttons and the Data-intensive Web', *New Media Society*, Vol. 15, No. 8, 2013: 1348-1365.

Gitelman, L. and Jackson, V. 'Introduction' in *'Raw Data' is an oxymoron*, edited by L. Gitelman. Cambridge: MIT Press, 2013.

Haes, A. de. 'Rijksoverheid zet alle cookies uit', *Webwereld*, 9 August 2012: http://webwereld.nl/nieuws/111422/rijksoverheid-zet-alle-cookies-uit.html.

Harvey, P., Reeves, M., and Ruppert, E. 'Anticipating Failure: Transparency devices and their effects', *Journal of Cultural Economy. Special Issue: The Device: The Social Life of Methods*, Vol. 6, No. 3, 2013: 294-312

Helmond, A. 'Trackers gebruikt op de websites van Nederlandse politieke partijen in kaart gebracht', *annehelmond.nl*, 11 June 2012: http://www.annehelmond.nl/2012/06/11/trackers-gebruikt-op-de-websites-van-nederlandse-politieke-partijen-in-kaart-gebracht/.

Hoofnagle, C. J. et al. 'Behavioral Advertising: The Offer You Cannot Refuse', *6 Harvard Law & Policy Review 273; UC Berkeley Public Law Research Paper No. 2137601*, 28 August 2012, available at SSRN: http://ssrn.com/abstract=2137601.

King, N. J. and Jessen, P. W. 'Profiling the mobile customer – Is industry self-regulation adequate to protect consumer privacy when behavioural advertisers target mobile phones? – Part II', *Computer Law & Security Review*, Vol. 26, No. 6, 2010: 595-612.

Koot, M. 'A Survey of Privacy & Security Decreasing Third-Party Content on Dutch Websites', 26 October 2012, available at http://www.madison-gurkha.com/press/2012-10-SurveyOfThirdPartyContent.pdf.

Latour, B. *Politics of nature: How to bring the sciences into democracy*, translated by C. Porter. Cambridge: Harvard University Press, 2004.

_____. *Reassembling the social: An introduction to actor-network-theory*. New York: Oxford University Press, 2005.

Law, J. and Urry, J. 'Enacting the social', *Economy and Society*, Vol. 33, No. 3, 2004: 390-410.

Leistert, O. 'Smell the fish: Digital Disneyland and the right to oblivion', *First Monday*, Vol. 18, No. 3, March 2013.

Lury, C. and Wakeford, N. *Inventive methods: The happening of the social*. Oxon: Routledge, 2014 (orig. in 2012).

Marres, N. 'Redistributing problems of participation', in *Material participation: Technology, the environment and everyday publics*. London: Palgrave Macmillan, 2012a.

_____. 'The redistribution of methods: on intervention in digital social research, broadly conceived', *The Sociological Review*, Vol. 60, 2012b: 139-165.

Marres, M. and Weltevrede, E. 'Scraping the Social? Issues in real-time social research', *Journal of Cultural Economy*, Vol. 6, No. 3, 2013: 313-335.

McDonald, A. M. and Cranor, L. F. 'The cost of reading privacy policies', *I/S: A Journal of Law and Policy for the Information Society*, Vol. 4, No. 3, 2008: 540-565.

McStay, A. 'I Consent: An Analysis of the Cookie Directive and Its Implications for UK Behavioral Advertising', *New Media & Society*, Vol. 14, No. 4, 2013: 596-611.

Raley, R. 'Dataveillance and Countervailence' in *'Raw Data' is an oxymoron*, edited by L. Gitelman. Cambridge: The MIT Press, 2013.

Rogers, R. 'Consumer technology after surveillance theory' in *Mind the screen: Media concepts according to Thomas Elsaesser*, edited by J. Kooijman et al. Amsterdam: Amsterdam University Press, 2008: 288-296.

_____. *The end of the virtual: Digital methods*. Amsterdam: Amsterdam University Press, 2009: 1-38.

_____. *Digital methods*. Cambridge: MIT Press, 2013.

Rogers, R., Weltevrede, E., Niederer, S., and Borra, E. K. 'National Web Studies: The Case of Iran Online' in *A companion to new media dynamics*, edited by J. Hartley, A. Bruns, and J. Burgess. Oxford: Blackwell, 2013: 142-166.

Ruppert, E., Law, J., and Savage, M. 'Reassembling Social Science Methods: The Challenge of Digital Devices', *Theory, Culture & Society*, Vol. 30, No. 4, July 2013: 22-46.

Savage, M. and Burrows, R. 'The Coming Crisis of Empirical Sociology', *Sociology*, Vol. 41, No. 5, 2007: 885-899.

Soltani, A., Peterson, A., and Gellman, B. 'NSA uses Google cookies to pinpoint targets for hacking', *The Washington Post*, 10 December 2013.

Tran, M. et al. 'Tracking the Trackers: Fast and Scalable Dynamic Analysis of Web Content for Privacy Violations', *Proceedings of the 10th international conference on Applied Cryptography and Network Security*, Singapore, June 2012.

Weltevrede, E. 'Introduction: Device-driven research' (working paper) in *Re-purposing digital methods: Exploring the research affordances of platforms and engines* (diss). Digital Methods Initiative, Department of Media Studies, University of Amsterdam, n.d.

Winter, B. de. 'Overheidswebsites sturen gegevens door naar derden', *nu.ul*, 1 November 2012: http://www.nu.nl/internet/2947863/overheidssites-sturen-gegevens-derden.html.

Wokke, A. 'OPTA waarschuwt overheidswebsites voor overtreden cookiewet', *Tweakers.net*, 6 September 2012.

Zuiderveen Borgesius, F. 'Behavioral Targeting: A European Legal Perspective', *IEEE Security & Privacy*, Vol. 11, No. 1, 2013: 82-85.

_____. 'Online Audience Buying', *Unlike Utechnolos Conference*, Institute of Network Cultures 8-10 March 2012, Amsterdam, 9 March 2012. Conference video available at http://vimeo.com/38840197.

[1] Rogers 2009. The Tracker Tracker tool in particular was developed in a collaborative project by Yngvil Beyer, Erik Borra, Carolin Gerlitz, Anne Helmond, Koen Martens, Simeona Petkova, JC Plantin, Bernhard Rieder, Esther Weltevrede, and Lonneke van der Velden during the Digital Methods Winter School 2012, 'Interfaces for the Cloud'. Project page: https://wiki.digitalmethods.net/Dmi/DmiWinterSchool2012TrackingTheTrackers.

[2] Savage & Burrows 2007; Marres 2012; Ruppert et al. 2013.

[3] Marres & Weltevrede 2013, p. 319.

[4] Marres 2012; Borra & Rieder 2014. Weltevrede n.d.

[5] Marres 2012b.

[6] Gitelman & Jackson 2013, p. 3.

[7] Ibid.

[8] Ruppert et al. 2013, p. 35.

[9] Law & Urry 2004; Callon & Muniesa 2005.

[10] Ruppert et al. 2013, p. 32.

[11] Gitelman & Jackson 2013, p. 4.

[12] Marres & Weltervrede 2013.

[13] McStay 2013, p. 597.

[14] Zuiderveen Borgesius 2013.

[15] Raley 2013.

[16] Zuiderveen Borgesius 2013; Leistert 2013.

[17] McDonald & Cranor 2008, p. 541; Zuiderveen Borgesius 2013.

[18] Raley 2013; Van den Berg & Van der Hof 2012.

[19] Ghostery. 'How It Works'. http://www.ghostery.com/how-it-works (accessed on 20 January 2014).

[20] In a later version Ghostery updated the 'Tracker' to Beacon (B) to prevent confusion with the general term Tracker.

[21] Evidon. 'The Evidon Blog'. http://blog.evidon.com/tag/ghostery/ (accessed on 7 March 2013).

[22] Evidon. 'Analytics'. http://www.evidon.com/analytics (accessed on 25 January 2014).

[23] 'What does Evidon do with Ghostrank information', https://www.ghostery.com/faq#q17 (accessed on 3 April 2014).

[24] Evidon. 'Better Advertising Acquires Ghostery'. http://www.evidon.com/blog/better-advertising-acquires-ghostery (accessed on 30 January 2014).

[25] Ghostery. 'Frequently Asked Questions'. https://www.ghostery.com/faq (accessed on 25 January 2014).

[26] Latour 2004.

[27] Harvey et al. 2013.

[28] Marres 2012a.

[29] Rogers 2013; Rogers et al. 2013.

[30] Latour 2005.

[31] See the Tracker Tracker project page (https://wiki.digitalmethods.net/Dmi/DmiWinterSchool2012TrackingTheTrackers) and the work of Helmond 2012 on Dutch political party websites.

[32] Gerlitz & Helmond 2013, p. 1349.

[33] Callon & Muniesa 2005.

[34] Because the law is formulated in a broad manner it applies to more tracking technologies than just cookies. 'Telecommunicatiewet, Artikel 11.7a', available at http://wetten.overheid.nl/BWBR0009950/Hoofdstuk11/i111/Artikel117a/geldigheidsdatum_03-09-2012.

[35] de Haes 2012.

[36] According to the 'Whois' information the domain names listed in the register are not all legally 'owned' by the government. Still, the government presents this list as their responsibility. The Website Register can be found at: http://www.rijksoverheid.nl/onderwerpen/overheidscommunicatie/eisen-aan-websites-rijksoverheid/websiteregister-rijksoverheid (accessed on 25 January 2013).

[37] The Dutch news site nu.nl paid attention to the study. See de Winter 2012.

[38] It entailed cleaning data and preparing the URLs before even using the tool and going through many error reports. More background to the method can be found at https://wiki.digitalmethods.net/Dmi/ThirdPartyDiary.

[39] Leistert 2013.

[40] Hoofnagle et al. 2012.

[41] Koot 2012.

[42] Brock 2010.

[43] Wokke 2012.

[44] For instance, in September raadvoordekinderbescherming.nl, which was redirecting to kinderbescherming.nl, was excluded in the October update. Therefore third party elements that were previously counted twice were counted only one time in October.

[45] The Digital Methods Wiki provides instructions for how to visualise Tracker Tracker data: https://wiki.digitalmethods.net/Dmi/WorkshopTrackingtheTrackers#A_42DMI_Projects_using_the_Track_the_Trackers_tool:_42.

[46] Checking the Whois and trace route of the IP address suggests that minjus.sdc.imetrix.nl was physically located in Amsterdam at the hosting company hostingbedrijf Redbee.nl.

[47] Leistert 2013.

[48] Koot 2012. See also Tran et al.

[49] This concerns the 'PREF-cookie', which also comes with Google Analytics. See Soltani & Peterson & Gellman 2013.

[50] Rogers 2008. This can be due to terms of services being non-negotiable (King & Jessen 2010).

[51] Elmer 2004, p. 130.

[52] Rogers 2008.

[53] Lury & Wakeford 2012, p. 13.

Comments are closed.

← Martin & Álvarez López to edit video essay section
De Cuir curates ICA Artists' Film Biennial →

Search

To search type and hit e

Share

1

Twitter

- #Student #protests at the University of #Amsterdam http://t.co/dh2Wm2BUgz 03:28:30 PM March 04, 2015
- NECS (@NECS_Network) 2015 notifications sent http://t.co/HWrdKO2kNN 02:58:40 PM March 04, 2015
- RT @necs2015: Acceptance/rejection notice was sent last Monday (March 2nd). In case you did not receive it, please check your spam folder. … 02:48:22 PM March 04, 2015

Follow @necsus_ejms 665 followers

NECSUS Newsletter sign-up

Email Address : Sign-up

Tag Cloud

alternative Amsterdam art Belgrade body book review British cinema colour conference culture digital documentary East-Central Europe ecology editorial exhibition festival film graduate green installation interview journal London media Milan NECS new media online open access phenomenology philosophy politics queer screen studies tangibility technology television touchscreen traces video war waste workshop

Subscriptions

Online

The electronic version of NECSUS_European Journal of Media Studies is published in Open Access and is therefore free and accessible to the public.

On paper

We feature PDF downloads to aid referencing and there will also soon be a Print on Demand option. Please consider the environmental costs of printing versus reading online.

Publisher

- Amsterdam University Press (AUP)

Partners

We would like to thank the following partners for their support:

- European Network for Cinema and Media Studies (NECS)
- Dutch Organisation of Scientific Research (NWO)
- Research School for Media Studies (RMeS)
- Further acknowledgements →

Editorial Board

Recent News

- Student protests at the University of Amsterdam
- NECS 2015 notifications sent
- 12th NECS Graduate Workshop in Rome
- Colour Fantastic @ EYE Film Institute
- Rancière in London

- Home
- Contact Us
- Disclaimer

Visual Network Analysis

Tommaso Venturini, Mathieu Jacomy, Débora Pereira

Introduction

In the last few years, a spectre has been haunting our academic and popular culture — the spectre of networks. Throughout social as well as natural sciences, more and more phenomena have come to be conceived as networks. Telecommunication networks, neural networks, social networks, epigenetic networks, ecological and economic networks[1], the very fabric of our existence seems to be made of lines and points.

Our fascination for networks is not unjustified and it is not new. Since Euler's walk on Königsberg's bridges[2], networks have proved to be powerful mathematical objects, capable of harnessing the most diverse situations where the connection of discrete elements is at stake. Yet, the recent fortune of networks derives less from their computational power than from their visual affordances. In the last years, the increasing availability of software for network manipulation has turned graphs into something that can be seen and manipulated. Turning graphs into maps and interface, this software has made network analysis available to more and more scholars particularly (but not exclusively) in the social sciences.

Yet the visualization of networks has so far lacked of reflexivity and formalization. Though all network analysis packages propose rich libraries of visualization functions, most literature on networks analysis is still centered on mathematical metrics[3] and does not detail how to *read* visualized network[4]. We painfully lack the conceptual tools to think about the projection of graphs in the space. The very vocabulary we use has been borrowed from mathematics (e.g. cluster, structural equivalence…) and geography (e.g.

[1] In the words of Fritjof Capra "The organic societies, like anthills and beehives, are metaphors that project the natural environment in the technological social space, as well as the structure of neurons and cells are models for understanding the networked world. In fact, networks naturally reflect the (dis)organization of the universe and nature." (1996)

[2] *Solutio problematis ad geometriam situs pertinentis*, 1736.

[3] The lack of interests of scholars working on graph mathematics for network visualization is not surprising. In solving to the problem of Königsberg's bridges, Euler performed the most classical of mathematics operations. He abstracted the formal structure of the problem from its empirical features: he took a city and turned it into a table of number. In doing so, Euler laid the foundation of discrete mathematics at the cost of separating the idea of network from its physical materializations.

[4] A notable exception can be found at the very beginning of the tradition of *social* networks analysis. Jacob Moreno, founder of this approach, was very explicit about the importance of visualization: "A process of charting has been devised by the sociometrists, the sociogram, which is more than merely a method of presentation. It is first of all a method of exploration" (1953, pp. 95-96). Though crucial for the founders of social network analysis, the reflection on network design progressively lost interest for their followers. Understandably fascinated by the parallel developments of graph mathematics, later social networks' analysts focused on statistics and progressively neglected networks design. On the history of social network visualization see Freeman, 2010.

centrality, bridging…) and need to be adapted to the new visual paradigm. This paper means to contribute to such reflection and propose a tentative framework for the visual analysis of networks.

To do so we will draw on the visual semiotics of Jacques Bertin (1967) and in particular on three of its variables: positions, size and hue. The papers will therefore be divided in three main sections, each addressing one of the three variables. Each section will explain how to project one variable on networks (using Gephi software as an example) and provide guidance on how to make sense of the resulting image. As position is, by far, the most important variable (for reasons that will be extensively explained), its discussion will occupy a largest part of the paper and will be divided in three sub-sections. To exemplifying our method of visual analysis, we will discuss a specific case study: a network of some 600 websites and hyperlinks related to the 2012 United Nations Conference on Sustainable Development of Rio de Janeiro (aka Rio+20)[5]. For each step in the analysis of this network, we will a) introduce the conceptual principle employed to read the network; b) exemplify the application of the principle on our case study; c) provide a tentative interpretation of the patterns observed on the network.

Visualizing node positions

How to give a position to nodes

Like geographical maps, graphs are generally two-dimensional representations, but unlike maps they cannot rely on a predefined set of projection rules. In a geographical representation, the space is defined a priori by the way the horizontal and vertical axes are constructed. Points are projected on such pre-existing space according to a set of rules that assign them a pair of coordinates and thereby a univocal position. The same is true for any Cartesian coordinate system, but not for network graphs. Nothing in network data predetermines where nodes have to be located in the graph. This has to do with the essentially discrete nature of graphs. Unlike geographical maps, graphs do not represent a continuous phenomenon (such as the distance between two landmarks), but a discrete one: two nodes are either connected or not. Therefore, as long as the edges are correctly drawn and link nodes that are connected in the dataset, nodes can assume whatever position without affecting the way the graph is read.

As a consequence, many different ways of positioning networks' node have been proposed through the years. In this article we will focus on a family of spatialization algorithms called "force-vector". Not only because these algorithms are, by far, the most commonly used in network spatialization, but they also because these algorithms have very interesting features. Force-vector algorithm work simulating a system of a physical forces: nodes are charged with a repulsive force that drives them apart, while edges work as

[5] 'Rio+20' is the short name for the United Nations Conference on Sustainable Development which took place in Rio de Janeiro, Brazil in June 2012 – twenty years after the landmark 1992 Earth Summit in Rio. At the Rio+20 Conference, world leaders, along with thousands of participants from the private sector, NGOs and other groups, came together to shape how we can reduce poverty, advance social equity and ensure environmental protection on an ever more crowded planet» (http://www.un.org/en/sustainablefuture/about.shtml).
The websites that compose the corpus that we will analyze have been selected according to two criteria:
1. if they are issued by organizations and groups active on environmental issues;
2. if they contain contents specifically related to Rio+20 or if they authors were present and active in the Conference.

springs bounding the nodes that they connect. Once the algorithm is launched it changes the disposition of nodes until reaching the equilibrium that guarantees the best balance of forces. Such equilibrium minimizes the number of lines crossings and thereby maximizes the legibility of the graph.

There is, however, a most interesting by-product of such visualization techniques: not only do force-vector algorithms minimize lines crossings, but they also give sense to the disposition of nodes in the space of the graph. In a spatialized network, spatial distance becomes meaningful: two nodes are close if they are directly connected or connected to the same set of nodes. Because of the very logic that drives them, force-vector algorithms assure that the distance among nodes is roughly proportional to their structural equivalence, that is to say the number of neighbors that they have in common (divided by the total number of their neighbors). Spatialization deliver an amazing result, it turns the discontinuous mathematics of graphs into a continuous space.

To spatialize our example network we used one of the many force-vector algorithms available in Gephi and called ForceAtlas2 (<add reference>) with the following parameters – LinLog mode, scaling 0.35, gravity 0.2, prevent overlap. Here is the result:

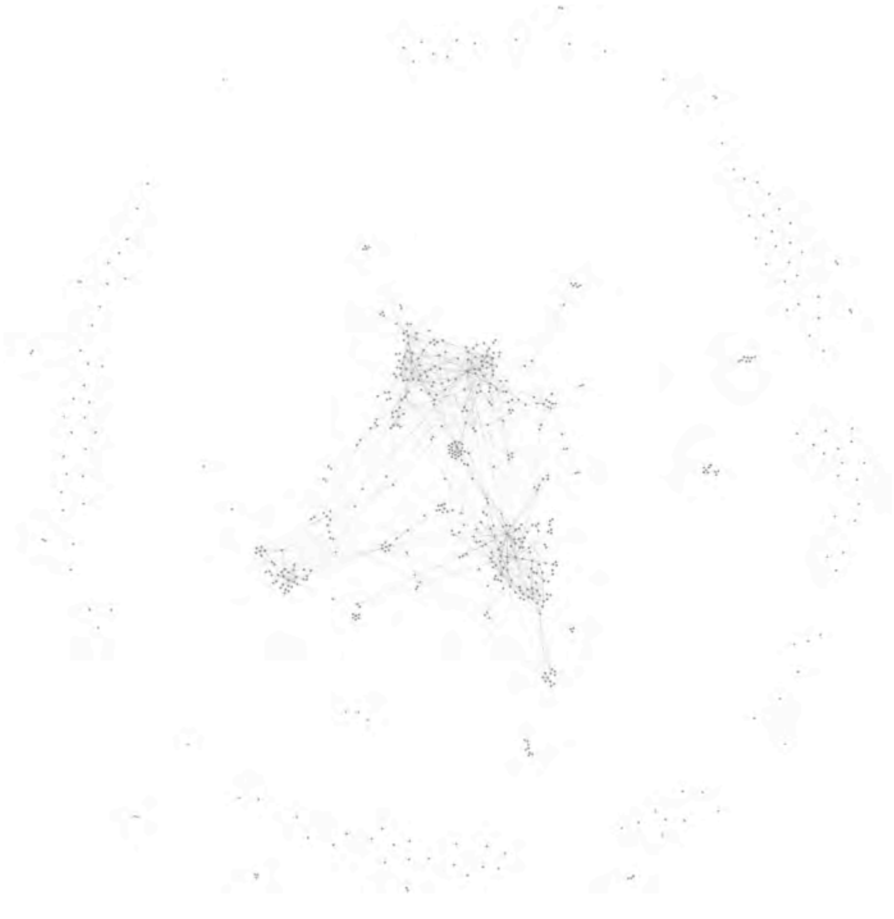


Figure. 1. The network after the spatialization. The main component is the most interesting part. The disconnected nodes form the ring (size and colors have not be modified).

How to interpret difference in density

Reading principle

In most networks, the spatialization reveals regions in which numerous nodes are assembled and regions that are empty or almost. These differences of densities (determined by the uneven distribution of the connectivity in the network) are revealed by the force-vector algorithm like different light exposures are revealed by chemical agents in photography. Spatialization generates visual patterns that translate the mathematical properties of the network. This translation is not free from distortions. Some properties are clearly visible, others are not. Some of the things that can be observed are meaningful, other are not. For example, the absolute position of nodes and cluster (at the top or bottom, left or right of the image) is completely arbitrary. What counts is the *relative* position of the nodes, their agglomeration and their separation. What matters is the clustering of the network.

To be sure, clusters could be detected in other ways. Andreas Noack, in particular, has shown that the mathematically mechanism of force-vectors corresponds to the computation of the clusters by modularity <citation Noack>, a technique often used to detect communities in networks <citation Newman>. Mathematical clustering however imposes a dissection of the network that is often too clear-cut. The advantage of visual techniques discussed in this article is that their fuzziness allows negotiating the frontiers of the clusters. These frontiers are naturally blurred, since clusters are not exclusive categories, but shades of density. Clusters may have clear boundaries, like cliffs separating a plateau from the valley, but most of the time their borders are gradual as the slopes of a mountain. The fuzziness of clusters' frontiers, by the way, is no obstacle to their recognition: a mountain is easy to see even is it impossible to say exactly where it starts and ends.

What is important is to be able to distinguish the clusters and to identify the empty zones between them. These zones are called "structural holes". The larger these holes are, the more they denote the absence of connection between the clusters. In dense graphs (such as those designed by hyperlinks' web or scientometrics networks), such absence is particularly significant and can be interpreted as a symptom of an opposition. Finally, we can remark that large clusters are often composed by smaller (and less distinct) sub-clusters. If large structural holes can be read as oppositions, smaller holes among sub-clusters may denote distinctions without opposition.

We can then summarize the reading principle of the first step by four questions:

- Which are the main clusters?
- Which are the main structural holes that separate them?
- Which are the sub-clusters within each cluster?
- Which are the smaller structural holes separating them?

Example

Which are the main clusters? In our example network it is easy to identify three main clusters at the top (A), at the bottom right (B) and at the bottom left (C). The clusters A and B are the largest and the easiest to identify. The cluster C is smaller and does not contain more nodes than the plurality of smaller sub-clusters scattered through the graph. The cluster C, however, is clearly distinguished from A and B and occupies *its own space*. This is why we count it as one of the main cluster of the network. The triangular shape of our network is thus the result of the three main clusters *pulling in three different directions*.

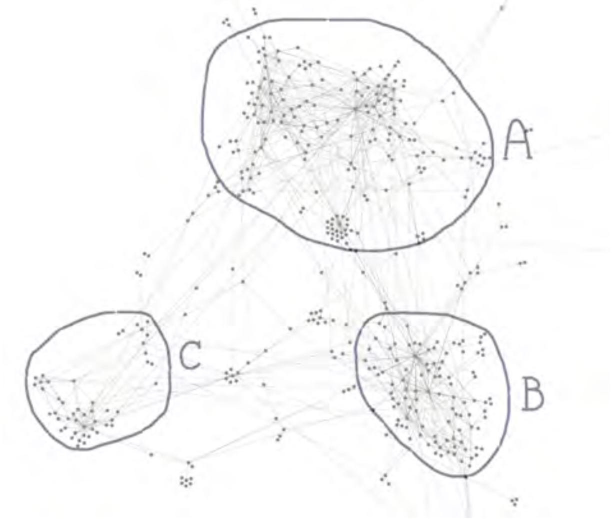


Figure. 2. The three main clusters (A-C)

Which are the sub-clusters? It is easier to distinguish the sub-clusters in the clusters A and B than in the cluster C that is significantly more compact. In A, we have identified two main sub-clusters a1 and a2 and three smaller groups of nodes. In C, we have identified three sub-clusters. In B, we decided not to separate any sub-clusters.

In the identification of the sub-clusters, there is always a part of subjectivity. Some sub-clusters are pretty evident (a1, a2 and c1), but most are not. Sub-clusters by definition smaller and less clear-cut than the main clusters and this can raise doubts on their existence: does c3 contains enough nodes to be interesting? Is a3 really separated from a1? We can leave these questions open for the moment. So far, the sub-clusters are only suggested to provide insights for the analysis.

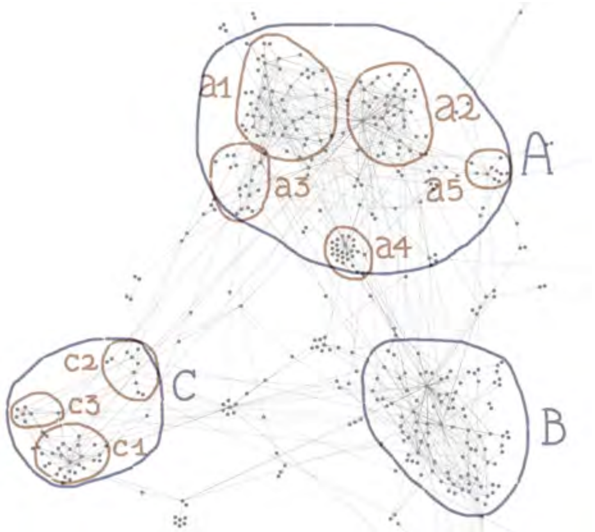


Figure. 3. The sub-clusters of A, B and C

Which are the main structural holes? In our network, there are four main structural holes: one at the center of the graph and three more separating the main cluster two by two. The cluster C is more isolated than the other two. The structural holes are very evident in this network: the absence of links between the main clusters is radical and demands to be explained.

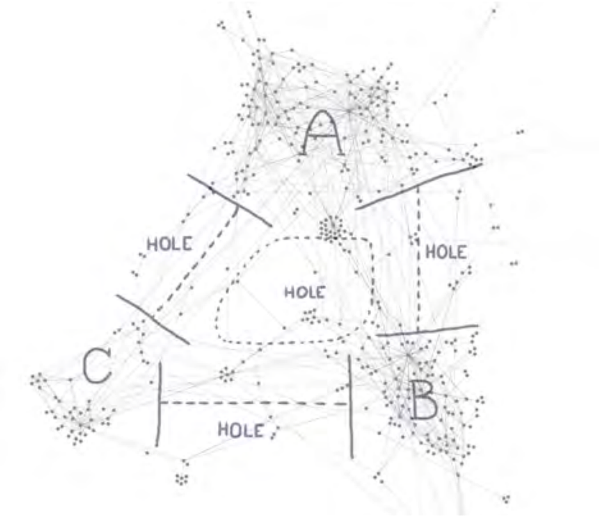


Figure. 4. The structural holes separating A, B and C.

Besides A, B and C, ten smaller clusters occupy different positions. These clusters can be divided in two groups: (1) The intermediary clusters M, E, F and L located among the main clusters. (2) The peripheral clusters K, G, I, D, J and H pushed towards the margins of the graph by the scarcity of edges that connects them to the three main clusters (some of them are so detached from the graph that we are led to consider the possibility of excluding them from the corpus).

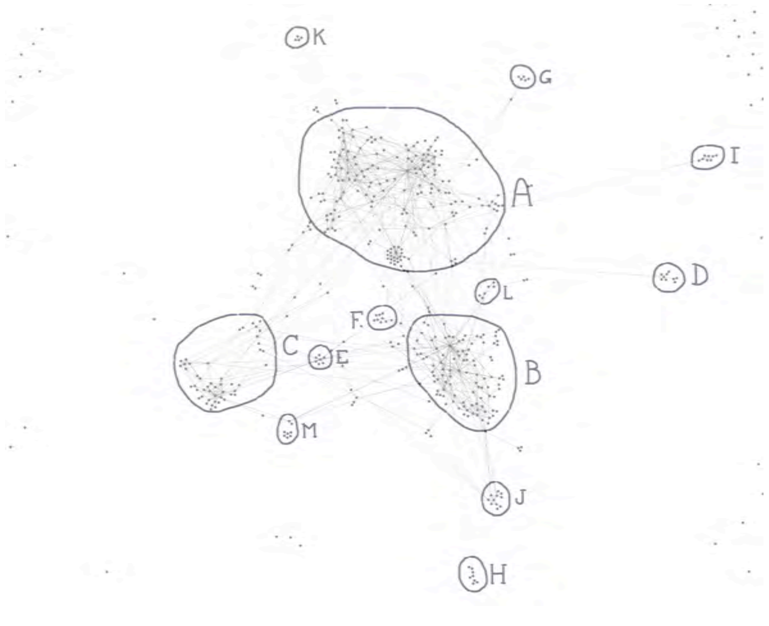


Figure. 5. Three main clusters (A-C) and ten smaller clusters (D to M)

Interpretation

After having spotted the clusters and sub-clusters of our network, we can try to make sense of them. The crucial aim of this phase is to find a suitable 'collective name' for each group of nodes. This is done by examining the nodes each cluster and trying to find what they have in common (at least what most have in common). Clearly some knowledge of the websites and their contents is necessary to discover similarities and that why visual network analysis should always be accompanied by some qualitative enquiry. In our case, all the websites had been visited and analyzed at the moment of the constitution of the corpus. Clustures constitute the main landmarks of the reading process and we will intensively refer to them in the next sections. The table below presents the main clusters and sub-clusters of our example networks:

Cluster	Actors	Contents
A **"NGOs and social movements"**	Social movements, environmental and human rights NGOs (mostly in Brazil).	Manifestations and social conflicts, indigenous issues, oppositions to dams, cultural events, courses agro-ecology, environmental education, forest management, 'Peoples Summit' event
A1 "marxist eco-socialism"	Main actors: Via Campesina, Movimento dos Sem Terra, Movimento dos Atingidos por Barragens	Ecological discourses inspired by the theories of the Marxist eco-socialism
A2 "environmental politics"	Actors active in southern and central highlands of Brazil	More heterogeneous than a1, its members do not exhibit the same militant politics, but a softer version of environmental politics, which does not engage in the struggle for human rights
A3 "Xingu river"	Dominated by Xingu Vivo movement, shows a connection between local entities in the north and northeast of Brazil and transnational NGOs, such as International Rivers and Conservation Strategy	Struggles against the construction for dams in the river Xingu
A4 "People's Summit"	Dedicate to the People's Summit, an event in Rio de Janeiro organized by social movements during the official United Nations Rio +20 summit	Protest against the official negotiations.
A5 "Brazilian government"	Centered around the Brazilian government, referred to by INRA in France, government of Bogotá and Forest Stewardship Council	
B **"international institutions"**	UN-related agencies and NGOs working on green economy and sustainable development. Main sites: ONU, official Rio+20, Unep	Reports on UN conferences and debates proposed. Recurring themes are alternative energy, clean water, carbon market, biomaterials and green ICT
C **"environmental and climate NGOs"**	NGOs for the preservation of forests, indigenous movements and scientific groups who advocate global warming as caused by humans.	Scientific articles, longer texts, campaigns and appeals for donation. Images of animals, forests, landscapes or nature, but without the presence of the man (often described as harmful to nature).
C1 "scientific websites"	Main actors: Real Climate blog, EcoEquity, Skepticalscience, Climateaudit, and Simondonner Indigenize	Debates on climate change and its causes
C2 "Mongabay"	Centered around Mongabay NGO	Information and pictures about nature and protest against its destruction
C3 "ecological Internet"	four websites (Forests, Rain Forests Portal, ClimateArk and Water Conservation) drawing contents from the Ecological Internet	Ecological Internet, self-define as a "non-profit organization that specializes in the use of the Internet to achieve conservation outcomes".

In general, the main clusters A, B and C form three coherent ensembles of websites. Despite internal differences exist, the websites in each of the main clusters are connected by hyperlinks because they share analogous interests and worries and a similar language. These specificities are also the reason of the separation between the three clusters. The NGOs in A and the institutions in B differ on every aspect: movements (A) VS. establishment (B); protest (A) VS. policy-making (B); mobilization (A) VS. planning (B). The institutions in B are also strongly opposed to the NGOs in C because of a deep difference in their values that opposes a pragmatic conception of modern societies and economies (B) to a radical questioning of the place of the humankind in the world (C). Finally, the websites in C and A are separated by the object they defend: social groups and communities for A and environment and ecosystems for C. A and C also employs different forms of engagement: the rationality of scientific knowledge and distant donation of money (C) against the emotions of social movements and first person participation (A). These differences explain why there are few bridges between our three main clusters. The actors tend not to link each, because of their ideological and practical opposition. They are more than different thematic clusters, they are opposed communities of interest.

How to interpret the size and density of clusters

Reading principle

Network clusters have two main properties that we can observe directly: their size and their density. Making sense of these properties is crucial to understand the balance of forces in the network.

The size of clusters is defined as the number of nodes they contain. The biggest clusters are the most visible on the Web and it is interesting to investigate the offline counterparts of such online significance. The Web is always a deforming prism. Minorities are sometime over-represented in online debate and some groups exist almost exclusively in the cyberspace <add citation to “Visibles mais peu nombreux”>. It is often relevant to compare the size of clusters on the Web to the presence of these actors in other public spaces (newspapers, institutions…).

The density of a cluster is a measure of its cohesion. Clusters are tight when they contain many edges and loose when few edges connect their nodes. In the case of the Web, a high number of in-cluster links may denote a the activity of a community: the actors know and acknowledge each other through their citations. A low density is also interesting. It may denote that the nodes do not know their neighbors or actively disregard them (because of competition or controversy). In low-density clusters, it is not the connections among their members that keeps them together, but the stronger separation with the rest of the network that *digs a ditch* of structural holes around them.

Example

The three main clusters count several dozens of nodes. Among them, A is the biggest and C the smallest. All the other clusters are definitely smaller with less than a dozen nodes.

As for the density, B is the only among the main clusters that is relatively dense. A and C are more spread out and clearly separated in smaller and denser sub-clusters (which appears to be more interesting than they larger parents). Cluster A is largely defined by a1 and a2. These two sub-clusters are large and dense and contains most of the nodes of A. a3 and a5, on the contrary, are scarcely dense and are distinguished from a1 and a2 only by their separated position. a4 is special case because of its star structure (see fig. 8).

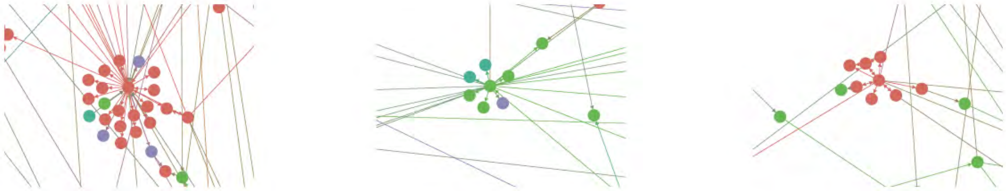


Figure. 6. Star structures: the sub-cluster a4 and the smaller clusters E and F

Similar little stars appears everywhere in the network (see fig. 8). They all are characterized by a central point surrounded by nodes that are only connected to the center but not among them. This is one case where the visual analysis of the network can be misleading: though they look compact, the stars are not particularly dense from a mathematical point of view. In the case of the Web, we will interpret these patterns as the symptom of the activity of an authority or a hub (more on this later on) and not as the sign of a communitarian activity. It is therefore important to distinguish stars from clusters: though they may look equally dense they are produced by very different mechanisms.

Cluster C contains a larger sub-cluster c1, a very sparse cluster c2 and another special case c3 that appears to be a clique. Cliques are group of nodes that are all connected to each other. It is rare to observe large cliques in natural networks, but it is possible to find small ones (with less than ten nodes). Quasi-cliques are similar structures where *almost* all the nodes are connected (see fig. 9). Cliques and quasi-cliques have no center, even if some nodes may appear more central than the others on the image (if all nodes are connected, no nodes is more connected than the others).

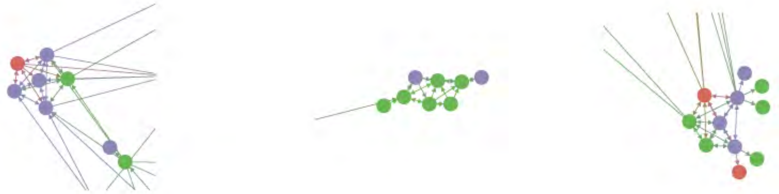


Figure. 7. Clique of the sub-cluster c3 and quasi-cliques of the clusters I and J

Interpretation

The **cluster A** “NGOs and social movements” is the largest of our network and this is probably due to the fact that it corresponds to an active community. The ‘occupation of the Web’ is an important issue for activists, both to assure their internal communication and to win the support of public opinion. The phenomenon of minorities’ over-representation may also be at play and this specific community may be particularly visible on the web. **Cluster C** “Environmental and climate NGOs” is also composed predominantly by NGOs and associations, but its lesser size may indicate a smaller or less active community. Finally, the **cluster B** “International institutions” is composed mainly by the numerous institutions gathered around the site of the United Nations (which explains the absence of sub-clusters). The links between institutions are often the simple mark of formal partnership. The sizes and shapes of the larger clusters seem therefore to be consistent with the different types of social organizations present in our network. The same can be said for the sub-clusters, which correspond to the divisions among the different types of associative groups present in A and C.

In the **sub-cluster a1** “Marxist eco-socialism”, a lot of information is circulated and discussed. This activity creates many links among the actors explaining the relatively high density. Blogs, in particular, play a central role in this group of sites (more on this later). The **sub-cluster a2** “Environmental politics” display a similarly intense communitarian activity (which explain the high number of links), but has a different thematic focus and is composed by a different type of actors, mostly NGOs. The separation between a1 and a2 may be in part explained by the fact that blogs tend to cite other blogs while NGOs prefer citing other NGOs. Though permeable, these two spheres remain relatively separated.

The **sub-cluster a4** “The People’s Summit” has the form of a star which can be explained by the fact that the social ecology scheme of the management of nature is well articulated, though it does not have the strengths nor the excessively referenced informational authorities. At the center of cluster C, the **sub-cluster c1** “Scientific debate on climate change” gathers an active and connected community. **c2** "Mongabay" and **c3** "Ecological Internet" represent two smaller groups of different types of actors (most NGOs) that though referring to the scientific blogs, are not confused with them. The clique structure of the **c3 sub-cluster** (experts in forest preservation, ecosystems and indigenous peoples) is explained by the fact that, with the exception of the sites Peoples Issues and New Earth Rising, all the other sites mirror the contents of the website Ecological Internet. Knowing the clique structure reinforces each actor notably in the search engine visibility, it would be interesting to investigate whether this strategy is deliberate.

How to detect centers and bridges

Reading principle

Now that we have identified the clusters, we can use them as landmarks to analyze two remarkable positions in spatialized networks: centers and bridges.

Centrality can be global (referred to the whole network) or local (referred to a single cluster). These two types of centrality are different. While the elements that are globally central are pulled in this position by the fact of being evenly linked to all the regions of the network, the element that are locally central tend to be linked predominantly within one cluster. Central positions (local or global) can be occupied by single nodes or by (sub-)clusters. In many cases, the center of the network (or the center of a large cluster composed by several sub-cluster) is just empty.

Bridges, on the other hand, are nodes or clusters that have connections with several clusters (two or more, but not all the clusters of the network). Bridges can be located outside the clusters they connect, if their connections are evenly distributed among them, or they can be located within one of the clusters, if they are more connected to it than to the others.

The following questions may help to identify systematically central and bridging elements:

- Which nodes or clusters (if any) are located in the center of the network?
- Which nodes or sub-clusters (if any) are located in the center of each cluster?
- Which nodes (if any) are located in the center of each sub-cluster?
- Which nodes or clusters (if any) are located among the main clusters?

- Which nodes (if any) are located among the main sub-clusters?

Example

Six nodes have been identified as centers of clusters and sub-clusters. Nature.com is the only node occupying the center of our network and the only node to connect the three main clusters. UN.org the website of the United Nations, is at the center of the cluster B. CupulaDosPovos.org.br is at the center of the sub-cluster a4 “People’s Summit” and RealClimate.org is at the center of c1 “scientific debate on climate change”. Finally, two nodes are central in smaller clusters IUCNWorldConservationCongress.org for E and Demilitarize.org for F. The presence of many edges around each of these nodes has helped us to detect their centrality.

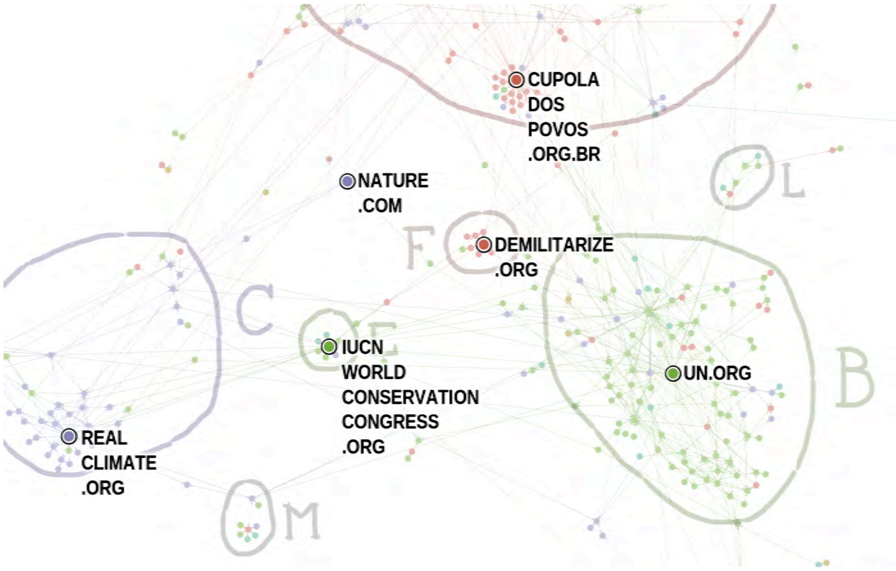


Figure. 8. Nodes in central position in the network or in their cluster.

Three clusters and seven nodes and are in bridge position. The clusters are the easiest to identify: E and F are located between B and C and L is located between A and B. Seven nodes are in a position to bridge between two clusters. Three nodes are simply ‘stretched’ between two clusters: GlobalVoicesOnline.com (A and C), NoGreenEconomy.org (A and B) and EffetsDeTerre.fr (B and C). Care2.com and IndianCountryTodayMediaNetwork.com *together* form a (3 links-long) bridge between A and C. Finally, Bndes.gov.br is an "internal" bridge located inside cluster C and connecting it to cluster B.

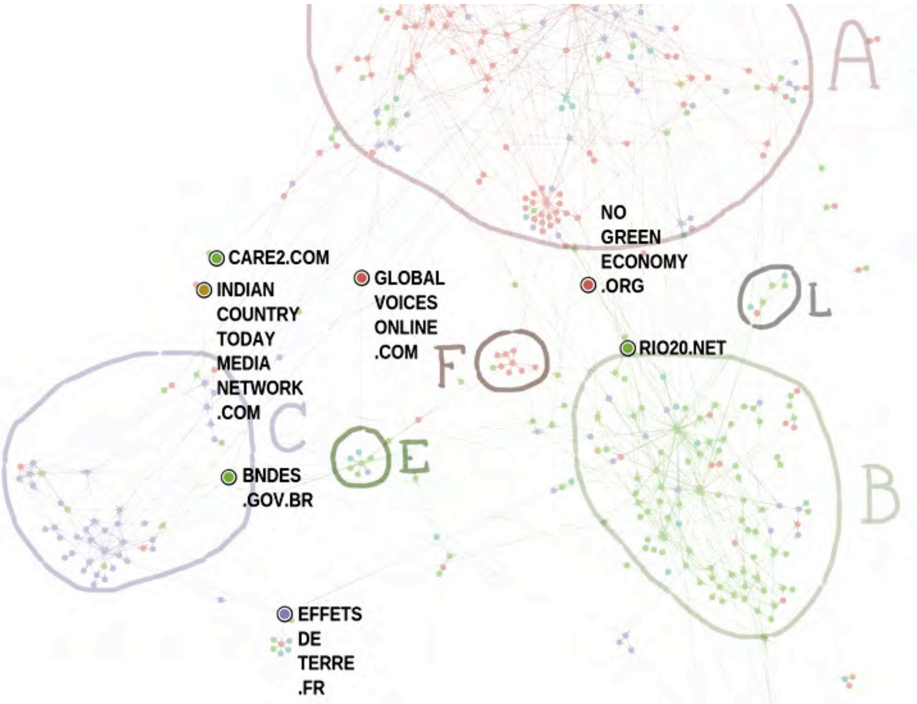


Figure. 9. Nodes and cluster in a bridging position.

Interpretation

Centers

Nature.com is the site of the famous scientific journal. It is an academic authority and is cited by all the many clusters, but not much (6 links in total). Of course, other websites exist that are cited by the nodes of the three main clusters (probably the major news websites), but they have not been included in the corpus because their content was not sufficiently focused on Rio +20 and its issues.

UN.org UN.org is a large portal linked to most of the numerous specialized institutions that constitute cluster B (and this explains its central position). The site contains static information on the United Nations: its mission, structure, Charter of Principles, the list of the member states and more. There is also a section of updates, daily news and highlighted dossiers.

RealClimate.org is a commentary website on climate science run by a group of renowned climate scientists and addressed to journalists and to the general public.

CupolaDosPovos.org.br is the center of the sub-cluster a4 “People’s Summit’ and its connections keep this star of nodes together. **IUCNWorldConservationCongress.org** and **Demilitarize.org** are in a similar position in clusters E and F.

Bridges

Interestingly, in our network, the role of bridge is played not only by nodes but also by small clusters. **Clusters E** “IUCN congress” is a bridge because of its focus on global warming mitigation. Mitigation is a key theme in conferences and events organized both by the institutions of the clusters B and the NGOs of the cluster C. As such it connects two distant regions of the network and facilitate their interaction.

Another site that promotes connections between clusters B “international institutions” and C “environmental and climate NGOs” is the blog **Effets de Terre** (an independent version of the blog that the journalist Denis Delbecq maintained from 2005 to 2007 in the French newspaper Libération on climate change and environmentalism).

Rio20.net present the program of the Cupola dos Povos event and is cited by several actors in clusters A and B. **NoGreenEconomy.org** is not an important website (the site seems ‘under construction’ and it contains only 5 posts). Its position between A and B is explained by the fact that the website is maintained by a group of activists, whose position are close to those of the social movements in A while criticizing punctually the approach of the institutions in B (and thereby citing them).

For all other cases, we have found no convincing interpretation. When the bridging position of a node cannot be confirmed by the qualitative analysis of its content, the best option is simply to ignore it. Unlike the larger patterns visible in the networks, single edges are not always significant. The aim of the visual analysis of networks is not to explain the position of each and every node, but to detect large trends and notable nodes.

Visualizing node sizes

How to give a size to nodes

We have now completed the part of the analysis based on the spatial position of nodes and we will start mobilizing the two other visual variables employed in the visual analysis of networks, starting from the size. In particular, we will now visualize the number of the edges arriving to or leaving from a node by changing the diameter of point that represent it. We will first change the size of the nodes according to number of incoming links (the in-degree) and then according to the number of their outgoing links (the out-degree). To do so, we have used the ranking palette of Gephi and set the diameter 1 for the smallest degree and 20 for the largest.

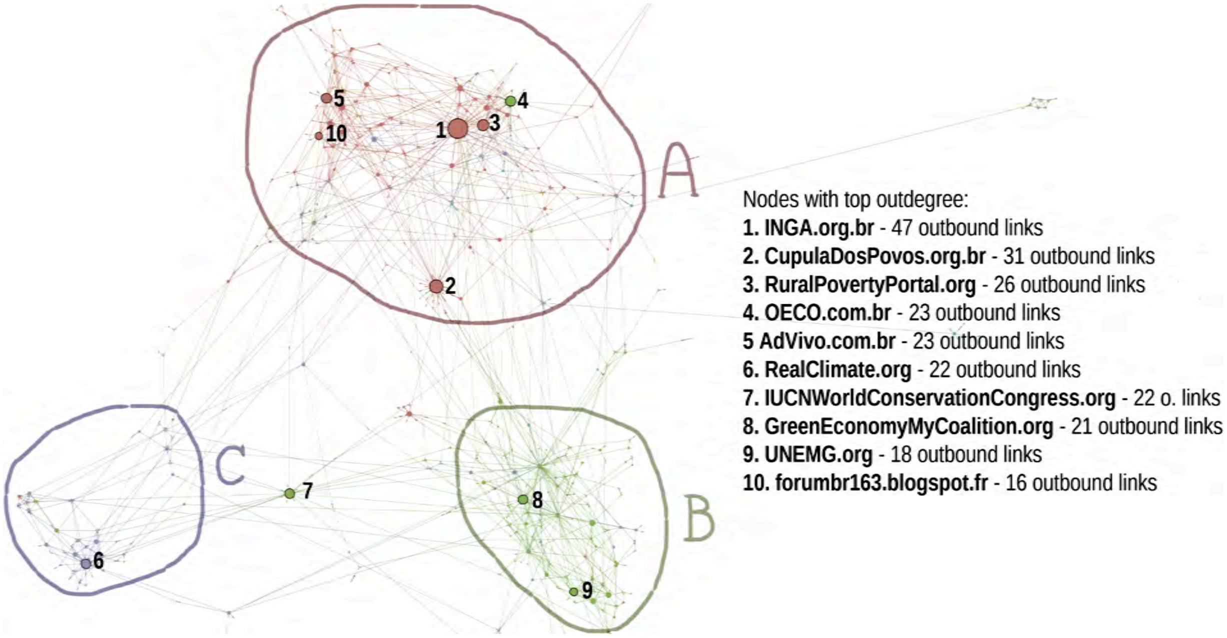
It is also possible (and indeed useful) to just look at the list of nodes sorted by their in-degree ou out-degree. Projecting the ranking on the spatialized networks, however, is also interesting as it allows identifying where the hubs and authorities are: are they central in a cluster or do they bridge different regions? Are they uniformly distributed or do they concentrate in some part of the graph? We could even go as far as to detect the local hubs and the authorities for each cluster (but we will not be so detailed in this article).

How to read the hierarchy of connectivity

Reading principle

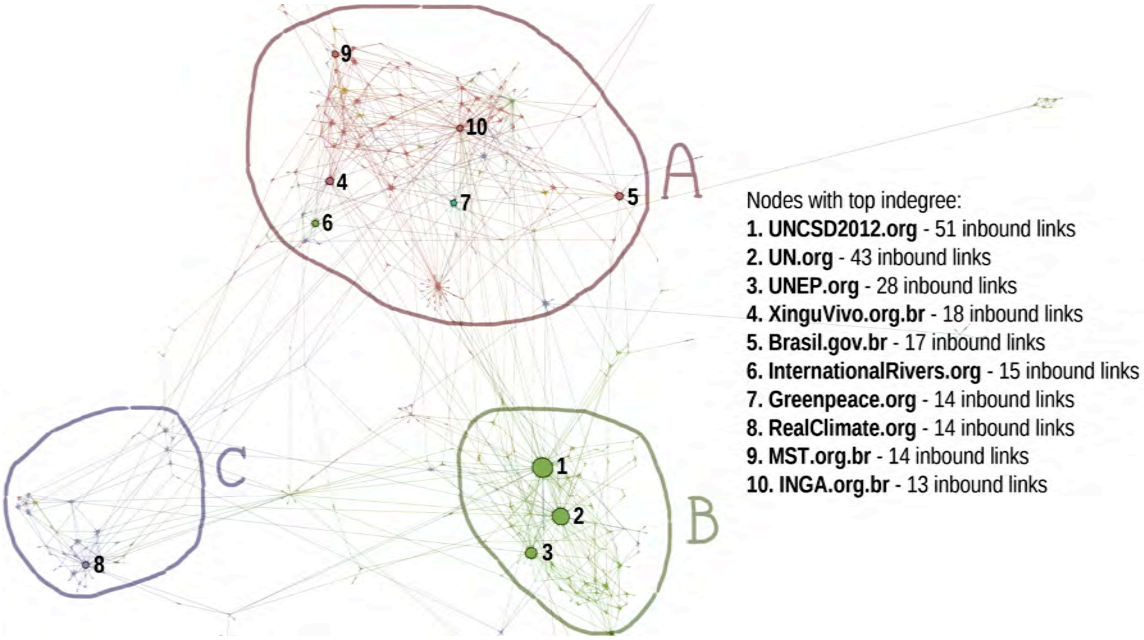
We will now consider the hierarchy of the most connected nodes. Following the tradition of network analysis, we will call ‘authorities’ the nodes that are the destination of many edges (inbound links) and ‘hub’ the nodes that are the origins of many edges (outbound links). Authorities are websites with a high visibility and toward which much of the traffic is addressed. Hubs are portals or websites that reference many other sites in the network. Both authorities and hubs tend to be influential nodes in the corpus.

Please remark that, in counting in-bound and out-bound links, we only take into consideration the connections within the corpus (and not all the hyperlinks that one website receives or sends). The website Nature.com, for instance, is certainly an authority in the World Wide Web, but it is not in our smaller network (despite its very central position).



Nodes with top outdegree:
1. **INGA.org.br** - 47 outbound links
2. **CupulaDosPovos.org.br** - 31 outbound links
3. **RuralPovertyPortal.org** - 26 outbound links
4. **OECO.com.br** - 23 outbound links
5 **AdVivo.com.br** - 23 outbound links
6. **RealClimate.org** - 22 outbound links
7. **IUCNWorldConservationCongress.org** - 22 o. links
8. **GreenEconomyMyCoalition.org** - 21 outbound links
9. **UNEMG.org** - 18 outbound links
10. **forumbr163.blogspot.fr** - 16 outbound links

Figure 10. Top10 authorities: the ten most cited sites in the corpus [top in-degree].



Nodes with top indegree:
1. **UNCSD2012.org** - 51 inbound links
2. **UN.org** - 43 inbound links
3. **UNEP.org** - 28 inbound links
4. **XinguVivo.org.br** - 18 inbound links
5. **Brasil.gov.br** - 17 inbound links
6. **InternationalRivers.org** - 15 inbound links
7. **Greenpeace.org** - 14 inbound links
8. **RealClimate.org** - 14 inbound links
9. **MST.org.br** - 14 inbound links
10. **INGA.org.br** - 13 inbound links

Figure 11. Top10 hubs: the ten sites citing the most other sites in the corpus [top out-degree]

Example of analysis

According to a power law often found in the Web (and in general in natural networks according to Barabasi, 2003), the distribution of the in-degree is very skewed in this graph, with the three most cited websites having 28 to 51 incoming links, while the rest of the top10 varies from 18 to 13. It is remarkable that all three main authorities of the graph are located in cluster B "international institutions". The rest of the top10 (except for one authority in cluster C) is located in cluster A "NGOs and social movement", whose high density of connection naturally produce local authorities.

As for the hubs, more than a half of the top10 (and the 5 biggest hubs) is located in the cluster A and again the density of such cluster can provide an explanation. It interesting to remark the presence of an important hub 'IUCN Congress' in a bridging position between B and C.

Interpretation

The main authorities of the network are all international institutions and this status seems to drive a large amount of hyperlinks to them. The three main authorities of the graph (uncsd2012.org, un.org and unep.org) are located in cluster B. The high density of this cluster and the lack of sub-clusters are largely due to the centripetal force of these three websites. These three websites are however local authorities: even if they receive links from other clusters, the largest part of their neighbors remains within cluster B.

Looking at outgoing links, given the high digital mobilization we observed in "NGOs and social movements", it is not surprising that most hubs are in cluster A and that these sites correspond to very active communities: INGA.org.br, RuralPovertyPortal.org and OECO.com.br are strongly engaged on rural ecology; AdVivo.com.br and ForumBr163.blogspot.fr on Marxist questions.

Visualizing node colors

How to apply a color to nodes

The last transformation we would like to operate on our network is to color its nodes according to the categories to which they belongs. This stage, of course, is only possible if the nodes have been categorized beforehand. To be sure, the same nodes can and (when possible) should be classified according to different systems of classification. Each classification system is project on the network as different layers of colors projected on the same background map. In our case, the nodes of the network had been categorized at the moment of the harvesting according to two different systems of classification: the approach to ecology that inspires them and the language in which they are written. Drawing on these classifications, we can use the partition panel in Gephi to attribute a different hue to each type of nodes.

It is important to remind that the color is a non-mixable visual variable. A node can be red or blue but not the two at the same time. When categorizing nodes, it is therefore necessary to employ exclusive categories. A website, for example, should be categorized as French or English, but not as both. If both languages are present on the same websites, researchers can add an additional category 'multi-lingual' (which is also exclusive).

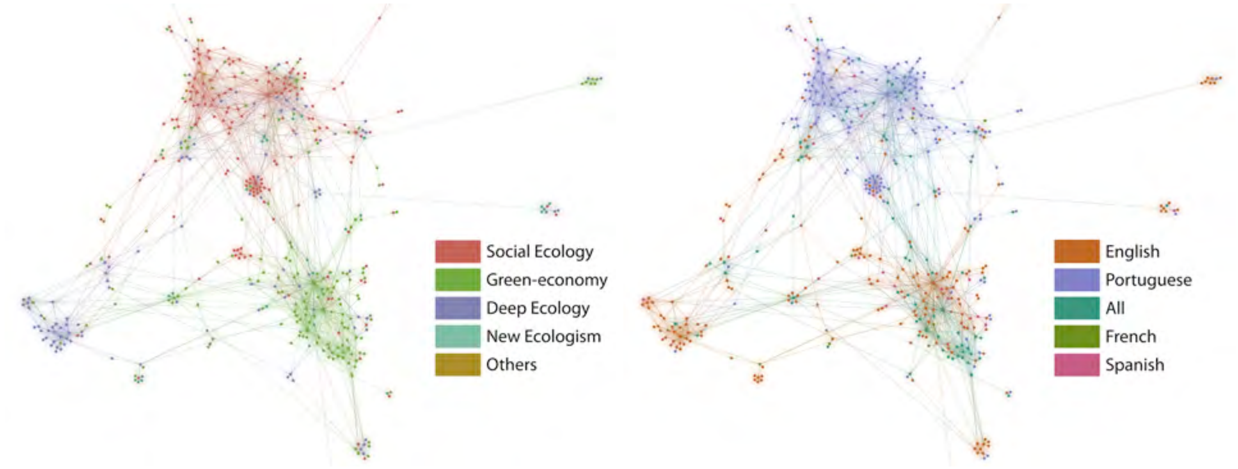


Figure. 12. The nodes of the network colored by approach and language

How to read the distribution of colored partitions

Reading principle

Having colored the nodes of our graph, we can now examine how the colors are distributed in the different regions of the network. In particular, it is interesting to observe if the nodes of the same color tend to be closer than nodes of different colors – creating a correspondence between the typology and the topology of the network. When such correspondence is observed it can be used as a basis to explain the patterns observed in the network.

Of course, the correspondence between categories and clusters is not always bijective: one category does not always correspond to one cluster. One category may colonize more than one cluster and two or more categories can associate to form a single cluster. Still, if the nodes of the same color tend to be closer than others, there is ground for interpretation. A interesting example of this situation is provided by the so-called 'hairball networks'. These are graphs that do no show any visible clusterization and are therefore difficult to analyze visually. However, when their nodes are colored, effects of polarization may appear. Even though the density of connection is homogenous all over the graph, nodes of different categories may still be visually separated.

Finally, when different layers of classification are present in the network, it is interesting to compare them and observe wether the different classifications produce the same borders in the network. Often this is not the case and sometime this explains why the correspondence between categories and clusters is not bijective.

Example

The most interesting differences between the websites of our network concern their different approaches to ecology. In particular, it is possible to find in our corpus websites corresponding to the three main 'schools' describe in the literature on ecology (Diegues, 2000; Koppes, 1989; Simonnet, 1979; Lipietz, 2012; Latour, 2004; Herber, 1964; Bramwell, 1989; Lynton, 1989; Lash Et Al, 1996; Zencey, 1989; Daly; Cobb, 1989):

- **Social Ecology**: explains environmental degradation as a result of capitalism and hierarchical division in society (between rich and poor, old and young, white, black and yellow). It advocates a return to primitive communitarian systems.
- **Deep Ecology**: deep ecology argues that nature was not given to humans, who have no right to use or exploit it. The objective of this type of ecology "is to preserve the nature of a hostile, essentially aggressive humanity" (Lipietz, 2012, p. 45).
- **New ecology**: emerged in the 60s, new ecology is directly opposed to consumerism.

An additional category, **Green Economy**, has been added to these to account for a large number of websites (32% of the corpus) that do not to fit in any of the previous categories and seem to be unified by the fact of proposing a synergy between ecology and market economy. Finally, as always, there are cases than cannot be pigeonholed in any category and are therefore classified as "**Others**".

Fig. 15 shows a clear correlation between our categorization and the topology of the network, as each of the three main clusters has a different dominant color. The cluster A "NGOs and social movements" is dominated by the social ecology approach; the cluster B "international institutions" is dominated by the green ecology; and the cluster C "environmental NGOs" by deep ecology. It is worth to remind that the spatialization algorithm we used do not take into consideration the categories of the nodes. The correspondence between hue and position is therefore a sign of strong correlation, so strong that we can make the hypothesis that the ideological agreement is a major driver of the connectivity in our network.

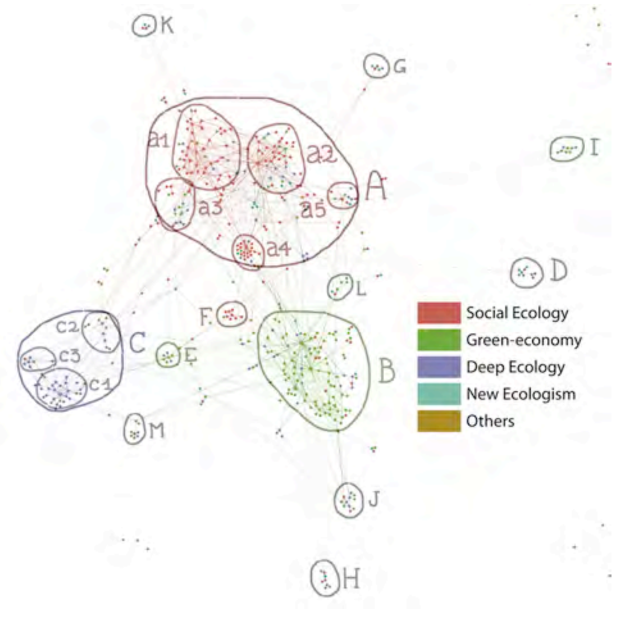


Figure. 13. All main clusters have a distinctive color.

Coloring the websites according to their language (fig. 16), we observe again a strong correspondence between typology and topology. Cluster A is largely composed by Portuguese websites, cluster B is divided between English and multilingual websites and cluster C is mostly in English. Since the proximity in the image indicates the (direct or indirect) connection, we observe a (not surprising) tendency to link websites of the same language.

It is interesting to observe the linguistic polarization of cluster B “international institutions", with English sites at the top and of multilingual sites at the bottom. Though clear-cut the linguistic separation in the cluster is not strong enough to produce a structural hole separating two different sub-clusters.

It also interesting to remark that cluster C “environmental NGOs” is also dominated by English websites and yet it does not merge to cluster B. Evidently, the organizational (NGOs VS international institutions) and ideological differences (deep ecology VS green economy) are, in this case, stronger than the linguistic bound.

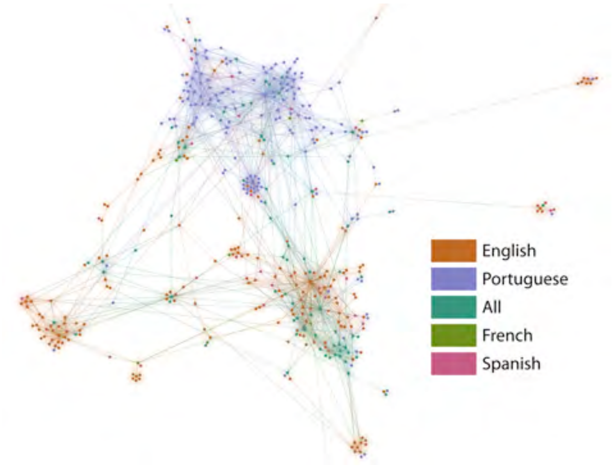


Figure. 14. Nodes colored by language

Interpretation

Drawing on the thematic and linguistic categorization, the separation between cluster A and B appears even deeper that we initially suspected. The two clusters are opposed by the type of organization that compose them (association VS institutions), by their approach to ecology (social ecology VS green economy) and by their language (Portuguese VS English and multilingual). This linguist difference seems also to imply a different geographical focus (local VS global). In this sense it is interesting to remark that the English pole of cluster B is closer to A (more connected) than the multilingual pole.

On the other hand, the structural hole between B and C cannot be explained by language and derives probably from the different positions in the debate. We can also observe that while the Portuguese websites cluster together the English websites do not. While Portuguese is the language shared by a community of local activists, English seems to be a neutral language used to address an international audience.

Conclusions

In this paper we have presented basic of the visual investigation of networks. This technique, we hope, will extend the ‘market’ of network analysis, by making the power of networks available to scholars with limited mathematical knowledge. By translating the key notions of graph mathematics (clustering, authority, bridging…) into the three visual variables of position, size and hue, we have tried to provide scholars with methods to analyze large and complex networks while sparing them most mathematical complications.

But there is more. Part of the interest of visual network analysis comes from the particular relationship between data and expertise that it proposes. Visual analysis entails a continuous iteration between the observation of the data and the interpretation of the findings. The continuous nature of two of the analytic variables (position and size) and the fact the third (color) depend on a manual categorization demand the constant engagement of the researcher's choice. Where are the limits of each cluster? Which nodes are central or more visible? Which are the bridges? Spatialized and ranked networks may suggest insights, but they never impose answers to these questions.

This has advantages and drawbacks. The main disadvantage of visual analysis is that it is impossible without some previous knowledge of the data and the phenomenon that they refer to. Without the help of Débora, who has constructed the hyperlink network and who has extensively studied Rio+20, there is no way we could have carried out such an insightful analysis. As every innovative research technique, visual research analysis is trapped in the "experimental regress" (Collins, 1975). Since both the method and its objects (the networks of hyperlinks, citations, words co-occurrence…) are still largely unexplored, it is hard to find a stable ground to establish their validity. How can we know that the patterns that we glimpse on the networks are not mere artifacts of the spatialization algorithm or projection of our previous knowledge? The only way out of these doubts is through the consistency between what we observe in the network and what we already know about the phenomenon it refers to. In our case, for example, we were comforted by finding a vast structural hole between social and deep ecology perfectly consistent with the long discussed difference between these two approaches. Also it was reassuring to find the websites of the organizers of the event around which the corpus was built (uncsd2012.org, un.org and unep.org) as the three largest authorities of the networks.

Our visual analysis, however, did not just confirm what we already knew about Rio+20 (little interest would have otherwise). It also offered a few notable surprises. The importance and separation of the green economy cluster was one of them, as well as the centrality of CupulaDosPovos.org.br and the bridging position of the site of its alternative summit, Rio20.net. We have already discussed these and other findings in the article and we will not come back to them in the conclusion. These examples serve only to illustrate the main advantage of visual network analysis. Precisely because it provides insights and not clear-cut answers, the visual investigation of network is primarily a method for exploratory analysis (Tukey, 1977). By encouraging scholars to engage with their networks (sometime to struggle with them), visual network analysis force researchers to assume an active attitude, to challenge and search ground for the previous knowledge and to open up to findings that they may not have thought to. Sometime visual network is frowningly compared to tasseography (the art of interpreting patterns in tea leaves, coffee grounds, or wine sediments). To a certain extent this comparison is not amiss: not unlike the best forms of divination, visual analysis is indeed meant to confront enquirer to their data, to explore their networks, to question their ideas. In this paper, we hope we have provided some guideline for it.

References

BARABÁSI, A. L. Linked. How Everything is Connected to Everything else and what it means for Business, Science and Everyday Life. Cambridge: Plume, 2003.

Bertin, J. (1967). *Sémiologie Graphique*. Paris - La Haye: Mouton.

BRAMWELL, Anna. **Ecology in the 20th Century**: a history. New York: Yale University Press, 1989.

Capra, F. (1996). *The Web of Life: A New Scientific Understanding of Living Systems*. *Colonial Waterbirds* (Vol. 20, p. 347). doi:10.2307/1521798

Collins, H. M. (1975). The Seven Sexes: A Study in the Sociology of a Phenomenon, or the Replication of Experiments in Physics. *Sociology*, 9, 205–224. doi:10.1177/003803857500900202

DIEGUES, Antonio Carlos Santana. **O mito moderno da natureza intocada.** São Paulo: Editora Hucitec, 2000.

DALY, Herman E.; COBB JR., John, B. **For the Common Good**: redirecting the Economy toward community, the environment and a sustainable future. Boston: Beacon, 1989.

Freeman, L. C. (2000). Visualizing Social Networks. *Journal of Social Structure*, 1(1).

HERBER, Lewis. **Ecology and Revolutionary Thought**. 1964. Disponível em: <http://ebookbrowse.com/gdoc.php?id=377184481&url=342a142f538180ceae5280ae1c3a84a3>. Acesso em: 3 jul. 2011.

KOPPES, C. 1989. "Efficiency, Equity, Esthetics; Shifting Themes in American Conservation". In: WORSTER, D. (ed.). **The Ends of the Earth: Perspectives on Modern Environmental History**. Cambridge: Cambridge University Press.

LATOUR, Bruno. **Politiques de la nature**: comment faire entre les sciences en démocratie. Paris: Éditions La Decouverte, 2004.

LASH, S.; SZERSZYNSKI, B.; WYNNE, B. **Risk, Environment & Modernity**: Towards a New Ecology. London: Sage, 1996.

LIPIETZ, Alain. **Qu'est-ce que l'ecolgie politique?** La grande transformation du XXI e siècle. Paris: Les Petits Matins, 2012.

LYNTON, Keith Caldwell. **International Environmental Policy**. Durham, N.C.: Duke University Press, 1990.

Moreno, J. (1953). *Who Shall Survive?* New York: Beacon House Inc.

SIMONNET, D. 1979. L'ecologisme. Paris: PUF (Que sais-je?)

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.

ZENCEY, Eric. Apocalipse now? Ecology and the perfidy of Doomsday visions. **Utne Reader**, n. 31, p. 90-93,1989.

Original Research Article

Big Data & Society
January–June 2016: 1–10
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/2053951716646135
bds.sagepub.com

How should we do the history of Big Data?

David Beer

Abstract

Taking its lead from Ian Hacking's article 'How should we do the history of statistics?', this article reflects on how we might develop a sociologically informed history of Big Data. It argues that within the history of social statistics we have a relatively well developed history of the material phenomenon of Big Data. Yet this article argues that we now need to take the concept of 'Big Data' seriously, there is a pressing need to explore the type of work that is being done by that concept. The article suggests a programme for work that explores the emergence of the concept of Big Data so as to track the institutional, organisational, political and everyday adoption of this term. It argues that the term Big Data has the effect of making-up data and, as such, is powerful in framing our understanding of those data and the possibilities that they afford.

Keywords
Big Data, Foucault, Ian Hacking, history, concept, discursive framing

Around 25 years ago, in a piece that was revised for publication in the highly influential collection *The Foucault Effect*, Ian Hacking (1991) asked the question 'how should we do the history of statistics?'. An apparently straightforward question that is likely to provoke some complex answers. What I would like to do here is to revisit that question in light of the emergence of 'Big Data'.[1] Put simply, I'd like to ask the question: how should we do the history of Big Data? Again this might seem straightforward, but by asking this question I am hinting at two things. First is the argument that we need to contextualise our understandings of Big Data within the history of social statistics. That is to say that we need to place Big Data within the genealogy of social data of various types. Second is the argument that we should approach this history by treating Big Data as both a material phenomenon and also a concept. Indeed, my central argument here is that we need to explore the *concept of Big Data* in historical, political and sociological terms. This is important because 'Big Data' is a concept that has achieved a profile and vitality that very few concepts attain. As such, its influence needs to be unpicked and understood. We need to understand the work that is being done by this powerful and prevalent concept.

To get things started though, what I am proposing here is that when thinking about Big Data we need to consider its history as being tied-up with particular ways of thinking. We then need to consider how this thinking is enacted in the development of certain infrastructures and in the industry of data analytics. This is to see Big Data as the entwinement of both a phenomenon and a concept. Big Data itself, with its capacity to track lives through archived and classified forms of individuated data, can be placed then within the genealogical lineage of the modern state (Beer, 2016). In this sense, it could be argued that we already have a history of Big Data that can be found in accounts of the history of the use of statistics to know and govern populations (see for example Desrosières, 1998; Foucault, 2007; Hacking 1990; MacKenzie, 1981; Porter, 1986, 1995; Elden, 2007). Similarly, we also have sociological resources that enable us to understand the power dynamics that reside within these accumulating data about people and populations (instructive examples here are Espeland and Sauder, 2007; Espeland and

Department of Sociology, University of York, Heslington, York, UK

Corresponding author:
David Beer, Department of Sociology, University of York, Heslington, York YO10 5DD, UK.
Email: david.beer@york.ac.uk

Stevens, 2008; and the various essays collected in Rottenburg et al., 2015). We even have a discussion of the use of Big Data in historical work – which asks how Big Data can be used by historians to gather or archive resources (Manning, 2013). And yet closer to the aims of this particular article, Halpern (2014), who focuses predominantly on the aesthetics of data, provides a historical account of the emergence of data since 1945. Despite all of this though we have something that is yet to be explored from a historical and sociological perspective, which is the work that is done by the very concept of 'Big Data'. That is to say that we have little understanding of the concept itself, where it came from, how it is used, what it is used for, how it lends authority, validates, justifies, and makes promises. In other words, we now need to work through a detailed account of what might be thought of as *the birth of Big Data*.

To reiterate, this stream of work is not concerned with the data itself, but with the discourse, terminology and rhetoric that surrounds it and which ushers and affords its incorporation into the social world. This is not to say that the specific material properties of Big Data are somehow unimportant, but rather that the way that these data are framed in particular rationalising discourses also needs to be treated carefully if we are to form a more detailed appreciation of the social implications of those data. It could be argued that in many ways the power dynamics of Big Data are to be found just as much in the way that those data are labelled and described as it is in the actual data themselves. Indeed, given the difficulties of data access and the technical and computing skills required to analyse Big Data, it might even be argued that the concept has far greater reach than the material phenomenon. However far the material consequences of Big Data might reach, the rationalities through which it is understood are likely to reach further. This article is dedicated to beginning to open-up this stream of work and is geared to developing a more contextual account of the concept of Big Data as it becomes embedded in organisational, political, social, cultural and everyday life.

The 'avalanche' of numbers

In Hacking's (1991) aforementioned essay on the history of statistics, which links into his other more substantial works on the same topic, he discusses some of the features of what he refers to as the 'avalanche of numbers'. This is where we can start to begin to contextualise the so-called Big Data revolution within a much longer history. Given the date of his piece it is obvious that when referring to this avalanche of numbers Hacking is not talking about the rise of digital technologies, smartphones, wearables or social media. Indeed, he is actually talking about an 'avalanche of numbers' that occurred around 1820 to 1840 as a new 'enthusiasm for numbers' (Hacking, 1991: 186) and a growing assemblage for data gathering took hold. Elsewhere this same period has been referred to as experiencing an 'explosion' of numbers. As Porter (1986: 11) observes, the 'great explosion of numbers that made the term statistics indispensable occurred during the 1820s and 1830s'. This 'explosion' or 'avalanche' of data occurred as 'nation-states classified, counted and tabulated their subjects anew' (Hacking, 1990: 2). In other words, the sense that we are being faced with a deluge of data about people is not something that is entirely new, in fact it has a long history. The type of data may have changed as might its analytics – with the shift toward commercial and algorithmic forms amongst other changes – but the lineage is clear. There are, of course, features of the current data moment that are in some ways novel but it is still interesting to note that this idea of a scaling up of social data, the feeling that we are facing an unfathomable flow of social data, itself has a history. The notion of an 'avalanche' gives the sense of the weight of escalating data resources that is comparable with the notion that data have suddenly got *big*. Both are based upon the feeling that there is a sudden and unstoppable wash of flowing data about people, a pooling of data that is on a scale that was not previously imagined. We have then both the phenomenon of the data interweaving with the way that it is imagined – the key difference here is that Hacking's terminology is based upon an *observation* about the increasing role of metrics in the 19th century whereas Big Data is a term that is commonly used in everyday discourse to refer to the data phenomenon of that very moment. Big Data is a concept which, like the data and methods associated with it, has a 'social life' (Savage, 2013). Referring back to the 19th century, Hacking (1991: 189) concludes that 'almost no domain of human enquiry is left untouched by the events that I call the avalanche of numbers, the erosion of determinism and the taming of chance'. We see here that already there was a sense that data harvesting was spreading out across the social world, and that this was accompanied by new means for analysing patterns in that data and for dealing with questions of probability.

By way of illustration let us turn again to Hacking's essay on how we should do the history of statistics. In that piece Hacking (1991: 191) points out that the 'the avalanche of numbers is at least in part the result of industrialization and the influx of people from the country to the town'. It was with the move towards industrialisation and the centralization of large parts of national populations in urban environments that the statistics about people began to escalate. The possibilities that came with these infrastructural changes were accompanied, Hacking argues, by a 'sheer fetishism for numbers' (1991: 192) and cultural shifts associated with the 'new countings' or 'new numberings' (1991: 191). This combination of social, technological and cultural changes led to the expansion of data that Hacking refers to as the 'avalanche of numbers'. As well as being counted in new ways, populations were also then ordered through categorisations. As Hacking (1991: 192) explains, 'when the avalanche of numbers began, classifications multiplied because this was the form of this new kind of discourse'. The emergence of new metrics also led to people being classified in new ways, which had powerful implications for how individuals and groups were perceived and treated. This new type of social ordering emerged with the need to manage the accumulating data about people. Hacking's (1991: 182) point here is that 'many of the modern categories by which we think about people and their activities were put in place by an attempt to collect numerical data'. Thus new categories emerged through which these new data might be gathered and by which they might then be analysed – leading to all kinds of 'classificatory struggles' (Tyler, 2015).

Part of the power associated with these escalating numbers, Hacking proposes, was to be found in their apparent objectivity. Statistical data, Hacking (1991: 184) claims, 'have a certain superficial neutrality'. It is this very appearance of neutrality that lends them an air of authority and which makes them so powerful. Tied in with this neutrality is the ability to use the numbers and categories to define what is seen to be normal and what is therefore seen to be abnormal. As Hacking (1991: 183) puts it, 'there are also statistical metaconcepts of which the most notable is "normalcy"'. Thus these accumulating data became a central means by which populations could be known and governed, and where understandings and expectations were produced alongside powerfully reinforced norms. Put simply, Hacking's (1991: 183) observation is that 'statistics of populations... form an integral part of the industrial state'. Industrial modernity brought with it expanded archives of data about populations (see Featherstone, 2000).

The result of all this is that statistics have become an important component in governance. Hacking's argument is that norms and classifications based around these types of data enable social facts to be brought into existence. Hacking (1991: 181) argues that:

> Statistics has helped determine the form of laws about society and the character of social facts. It has engendered concepts and classifications within the human sciences. Moreover the collection of statistics has created, at the least, a great bureaucratic machinery. It may think of itself as providing only information, but it is itself part of the technology of power in a modern state.

Statistics are then incorporated into the very infrastructures and modes of governance of the state – in the last twenty or so years we might also add corporate and commercial data gathering to this. The result is that the categories and modes of reasoning surrounding these data become part of the formal and legal structures of the state, with direct implications for how people are treated (for one example in relation to immigration see Schinkel, 2013). Again, in Hacking's (1991: 194) words, the 'bureaucracy of statistics imposes not just by creating administrative rulings but by determining classifications within which people must think of themselves and of the actions that are open to them'. These emerging numbers quickly came to define how people saw themselves, how they saw others and, complimenting these, how limits and boundaries were placed around actions and opportunity. So, Big Data can be placed within this long history of social statistics, but we might also note that there is undoubtedly an intensification in the scale of data over that time, particularly as commercial organisations have joined in with the state to increase the infrastructures, scope, accumulation and deployment of data (Ajana, 2013; Beer, 2016; Kitchin, 2014).

As this would suggest it is important to situate this article within the history of the development of statistics but this is beginning to take us into territory that resides beyond the remit of this article, what I'd like us to take from this is that the expansion of data and metrics is not something that can be isolated to a particular moment in the recent past. Rather the powerful ordering presence of data has been felt for some time, as has the sense that there is an overwhelming deluge of information associated with the march of modernity. What I would like to do here is to put this particular historical context to one side, there are other places we might go in order to explore the genealogical history of what is now referred to as Big Data (for an overview of the history of social statistics in relation to Big Data see chapter 2 in Beer, 2016). Instead, let us be aware of the infrastructural, technical and cultural history of such data whilst focusing our attention upon the relatively short life of the actual concept or term *Big Data*. We may need further work that delimits the particular material and ontological properties of the current form that this social data takes – such as the work being conducted by Kitchin (2014; see also Kitchin and McArdle, 2016) – but let us turn our attention elsewhere for the moment. The concept of Big Data has a short history which is part of a much longer series of developments stretching back hundreds of years. I want to use the remainder of this article though to argue that it is the work that is being done by this particular concept that requires our attention – particularly if we are to continue to attempt to develop a more complete and contextual understanding of the influence of data today. As such, the point is that the history of Big Data as a phenomenon can tracked back through the pages of those histories of social statistics – even if much more work is needed in order for a more global and 'connected' (Bhambra, 2014) history of statistics to be developed – but what we have little understanding of is the *birth* and *life* of Big Data as a concept. It is this concept that needs to be tracked, unpacked and examined. Indeed, the aim of this article is to begin to map-out a programme of work that needs to be completed in order for us to fully understand the politics of Big Data.

Treating 'Big Data' as a concept

This is where I would like us to prize ourselves away from the data themselves, to begin to think historically about how these data are conceptualised. It is by acknowledging the long history of the accumulation of data about individuals and populations that we can begin to make a departure into seeing the different ways that data are presented in conceptual terms – and thus where we might begin to see more clearly the importance of the project of exploring Big Data as an interweaving of a material phenomenon and circulating concept.

Both Ian Hacking and Stuart Elden suggest that the only way to really understand the power and influence of concepts is to see them in their historical context. Hacking's (1991: 184) position is that we need to explore 'the relationship between concepts in their historical site' (Hacking, 1991: 184). Similarly, Elden (2013a: 15) argues that 'conceptual history is important because of its emphasis on terminology, and the relation between meaning and designation; contextualist approaches are crucial in stressing the importance of reading texts within the situations in which they were written'. In relation to territory Elden's (2013a: 15) position is that 'territory is a word, concept and practice; and the relations between these can only be grasped historically' – this is a project that Elden (2013b) expands upon in much greater detail in his book *The Birth of Territory*. The point here is that we can only understand certain social phenomena through their discursive and conceptual formulations, and we can only understand these conceptual formulations by thinking historically about them. Both Hacking and Elden place concepts at the centre of their historical analyses.

Hacking offers further explanation of his position by claiming that:

> the organization of our concepts, and the philosophical difficulties that arise from them, sometimes have to do with their historical origins. When there is a radical transformation of ideas, whether by evolution or by an abrupt mutation, I think that whatever made the transformation possible leaves its mark upon subsequent reasoning. (Hacking, 1991: 184)

Concepts are a product of their historical origins, we might conclude, but they then also have social reach and influence themselves. The organization of our concepts can then be at the heart of social transformations – the transformation of ideas is a powerful thing. These concepts and the transformations of which they are a part leave, Hacking suggests, an indelible mark on future reasoning. They leave their mark on the way that the social world is comprehended and acted upon. If we are to pursue the concept of Big Data with this in mind, then we would not just be looking at the concept for its influence during the lifetime of its use but also its potential influence on future reasoning. We would also need to look at the discursive frameworks and modes of reasoning that fed into the concept of Big Data. Thus a genealogy of a concept like Big Data aims to capture the emergence of a concept as a part of a historical lineage of reasoning that shoots out into the past and the future. It is a moment, but a moment in which we might reveal something longer term.

We can of course see the influence of Michel Foucault echoing through Hacking and Elden's approaches. We can amplify these echoes by turning to a relatively well-known interview with Foucault which was originally published in 1980. The interview focuses upon questions of method. Amongst various aspects of Foucault's approach discussed in that interview, a particular theme emerges concerning the role of concepts in shaping social realities. Here Foucault describes some of the methods he deployed in his works and focuses in particular upon the need to explore conceptual processes in the formation of the social world. He focuses upon his concern with understanding the different ways in which truth is produced through practice. As Foucault explains:

> To put the matter clearly: my problem is to see how men govern (themselves and others) by the production of truth (I repeat once again that by production of truth I mean not the production of true utterances, but the establishment of domains in which the practice of true and false can be made at once ordered and pertinent). (Foucault, 1991: 79)

It is unusual to find something so crucial hidden within brackets. Foucault is interested in exploring the ways in which truth is produced so as to see how those truths limit understandings, actions and practices. His intention is to use events and moments to open up these regimes of truth, and to understand how these regimes of truth activate practices. As he put it, 'eventalizing singular ensembles of practices, so as to make them graspable as different regimes of "jurisdication" and "veridiction": that, to put it in exceedingly barbarous terms, is what I would like to do' (Foucault, 1991: 79). His intention then was to grasp the practices that translate into the boundaries and limits of jurisdictions and, alongside this, to see how truth is verified in different ways – it is notable that he argues elsewhere that markets are the 'sites of veridiction' (see Foucault, 2008: 32). In doing this, his aim is to 'resituate the production of true and false at the heart of historical analysis and political critique' (Foucault, 1991: 79). Elsewhere Foucault (2014: 7) describes this production or manifestation of certain regimes of truth as a process of 'alethurgy' – which is concerned with understanding the 'manifestation of truth' as central to the formation of power structures. These regimes of truth and their limited powers can then be seen to be found in the discourse surrounding certain practices.

At this point in the interview Foucault's attention shifts to the notion of 'programmes' in order to exemplify and explain these wider objectives. It could be read that when he talks of programmes he is talking about the set of practices in which regimes of truth are imagined and then made possible. He talks here of programmes of activity that are not always realised, but which can be used to explore how ideas are projected onto the social world. In the interview, Foucault is questioned on the separation of these programmes from the reality of what is happening on the ground. Foucault's response is to emphasize the importance of understanding how the world is imagined in order to understand how it unfolds. As he explains:

> Bentham's *Panopticon* isn't a very good description of 'real life' in nineteenth-century prisons. To this I would reply: if I had wanted to describe 'real life' in the prisons, I wouldn't indeed have gone to Bentham. But the fact that this real life isn't the same thing as theoreticians' schemas doesn't entail that these schemas are therefore utopian, imaginary, etc. One could only think that if one had a very impoverished notion of the real. For one thing, the elaboration of these schemas corresponds to a whole series of diverse practices and strategies. (Foucault, 1991: 81)

Obviously referring back to the work he did for his 1975 book *Discipline and Punish*, Foucault is arguing that the types of programmes or imagined possibilities captured in concepts like the panopticon are important. These types of concepts become woven into reality in different ways, they become part of practice as they are cemented into jurisdictions, boundaries and as they verify, authorise and select what comes to be. Concepts, or programmes, can then be elaborated in practice in ways that are not always obvious. Separating them from reality would be a mistake. Thus we cannot see Big Data as being a programme that exists outside of the practices of the use of data on the ground. Similarly, Big Data may not necessarily be a very good concept for seeing the 'reality' of everyday life, but it is a good concept for understanding how visions of contemporary data are incorporated into the imagining of life, the production of truths and the liminal work that contains the social world. Big Data is undoubtedly a part of contemporary strategies and practices. The point here is that Big Data may be treated as a programme of thought that needs to be analysed in this way.

Foucault extends this point further. He adds that these 'programmes induce a whole series of effects in the real (which isn't of course the same as saying that they take the place of the real): they crystallize into institutions, they inform individual behaviour, they act as grids for the perception and evaluation of things' (Foucault, 1991: 81). As such, imagined programmes and conceptual formations translate into regimes of truth. That is to say that they solidify into practices, organisations, institutions and behaviours. To apply this to Big Data we could imagine how this concept carries with it 'grids for the perception and evaluation of things'. That is to say, that it is not just the evaluations that come from the applications of Big Data themselves, but that the concept of Big Data as a programmatic mode of reasoning also brings with it the values and norms that provide the means for evaluating and judging. It is not just the data that afford judgments, it is also the very concept of Big Data itself that shapes decisions, judgments and notions of value – as it brings with it a vision for particular types of calculative or numerical knowing about individuals, groups and the social world. These are legitimised in the case of Big Data by notions of its scale and the eradication of error and inefficiency (for a discussion of scale and accuracy in Big Data see boyd and Crawford, 2012). These programmes of Big Data arrive with a thirsty desire to render measurable.

For Foucault, whether or not these imagined programmes are ever fully realised is not necessarily important, rather it is the influence that those imagined programmes have in shaping practice. It is also the broader rationality that they reflect. In the case of Big Data we might not see the project or its imagined potential realised in full, but the concept has already been influential far beyond the reach of the data in many respects. Just because programmes are never fully realised does not mean that they are somehow insignificant, especially when they achieve the prominence that is enjoyed by the Big Data movement. Rather we should see how that programme was pursued and how the imagined outcomes became part of practices and strategies – or how they relate or encapsulate a broader art of governance, political economy or prevalent forms of rationality and reasoning. According to Foucault:

> These programmings of behaviour, these regimes of jurisdiction and veridiction aren't abortive schemas for the creation of reality. They are fragments of reality which induce such particular effects in the real as the distinction between true and false implicit in the ways men 'direct', 'govern', and 'conduct' themselves and others. (Foucault, 1991: 82)

Here the *programme* shifts to being about *programming*, about setting up the codes of social life, with these conceptual framings being fragments of reality. Such programmes fracture into reality, for Foucault. It is in the job of unpicking these fragments that he is interested. Foucault's use of true and false might seem blunt, but he is pointing to the powerful ways in which such programmes set rigid limits. Such conceptual programmes, for Foucault then, 'induce effects' and make things happen. They are part of governance and they act to shape conduct by contributing towards these regimes of truth that inform behaviour. As he further explains, we need to attend to 'the correlative formation of domains and objects and . . . the verifiable, falsifiable discourses that bear on them; and it's not just their formation that interests me, but the effects in the real to which they are linked' (Foucault, 1991: 85). The challenge, once we take such schema as being important to the conduct of the reality of the social world, is in thinking about how to explore their emergence and effects.

Framing Big Data

Recently we have seen some attempts that begin to think through or suggest the need to think through the role of the concepts and discourses that surround data today. For example, Rob Kitchin has suggested that we need to the look at the political and economic framing of Big Data. He suggests that we should look at 'how a powerful set of rationalities is being developed to support the roll-out and adoption of Big Data technologies and solutions' (Kitchin, 2014: 126). Kitchin indicates that this is part of a broader project whilst focusing his discussion across four 'major tasks': 'governing people', 'managing organisations', 'leveraging value' and 'producing capital'.

For Kitchin these underpinning rationalities need to be explored because they play such a potent part in the integration of Big Data. These rationalities are to be found in the discursive regimes of Big Data, and thus these regimes need detailed and careful attention in order to understand the power dynamics of Big Data. One way into this is to look at the logic that is woven into the Big Data movement. As Kitchin (2014: 126) puts it:

> The power of the discursive regimes being constructed is illustrated by considering the counter-arguments – it is difficult to contend that being less insightful and wise, productive, competitive, efficient, effective, sustainable, secure, safe, and so on, is a desirable situation. If big data provide all of these benefits, the regime contends that it makes little sense not to pursue the development of big data systems.

Considering the connotations and implications of those powerful underpinning rationalities reveals the potency of the discourse here. Big Data brings with it a force to comply and a rationality that is hard to critique or resist. This can potentially be seen to have a kind of neoliberal reasoning or rationality at its core, one based upon the use of data as the mechanism by which the model of the market may be rolled out across the social world (for a discussion see Beer, 2016). As a result of these considerations Kitchin concludes that 'what is presently required, through specific case studies is a much more detailed mapping out and deconstruction of the unfolding discursive regimes being constructed' (Kitchin, 2014: 126). There are undoubtedly parallels here between Kitchin's suggestion and the project I'm mapping out in this article. Kitchin gives us a starting point through which we might channel the type of observations drawn out from Foucault, Elden and Hacking's work. It is pressing, as Kitchin has put it, that 'given the utility of the data, there is a critical need to engage with them from a philosophical and conceptual point of view' (Kitchin, 2014: 185). What we need then are the conceptual and historical resources that will enable us to develop a richer understanding of the discourses and rationalities of Big Data. It is this point that we have reached. We need now to work out ways of expanding such a set of insights and to flesh out this approach. My suggestion is that we focus our attention centrally upon the term of Big Data itself and begin to explore it historically and conceptually. This will provide a focal point for responding to Kitchin's more general call. Part of this will require us to not just challenge or dismiss but to carefully unpick 'boosterist discourses declaring their positive disruptive effects' (Kitchin, 2014: 192). It is this unfolding discursive regime that needs attention – by illuminating the rhetoric orbiting around the concept of Big Data. There is undoubtedly more to Big Data than its discursive framing, it has material properties that make it Big Data (see Kitchin and McArdle, 2016), yet the particularities of that discursive framing shape those material presences and the integration of Big Data into broader social structures and orders.

Elsewhere there are some other rare occasions in which the power of such discourses has been acknowledged. For library information scientist Ronald E Day this discourse centres around a particular set of claims. The shift for Day is away from notions of 'information', which implies something flexible and informed, and toward something much more rigid. According to Day (2014: 3), 'more recently, the discourse of "data", conceived as a form of auto-affective presence or "fact," has come to supersede the trope of "information"'. The shift then is towards the notion that data is equivalent to facts, and thus then away from a more open vision of information. This, for Day, is an important shift that makes contemporary notions of data much more powerful in social formations. He continues this line of argument by claiming that these:

> claims for knowledge are presented as immediate – 'factual' – rather than as emergent through technologies, techniques, and methods, on the one hand, and interpreted through theory or a priori concepts, on the other hand. *The data says...; the data shows us...; we are only interested in data (not justifications/excuses/your opinion/your experience)...; big data and its mining and visualizations gives us a macroscopic view to see the world anew now* – these and similar phrases and tropes now fill the air with what is claimed to be a new form of knowledge and a new tool for governance that are superior to all others, past and present. (Day, 2014: 134; italics in the original)

This presentation of data as facts is crucial, for Day, in understanding the powerful role played by those data. Again, as with Kitchin, Day explores how the data are presented in compelling and even irresistible ways. In the above passage Day offers some illustrations of how this type of discursive framing works in practice. In these formulations the data is seen to be objective, neutral and telling – it is not something to be questioned or interrogated, it is rather a social fact around which behaviour should be bent. It is seen to be a tool for governance that cannot be questioned or rivalled with subjective opinions. Data is seen, in this formation, to be unquestionable, accurate and over-arching in its panoramic view of the social world.

These positions provide some revealing opening insights, but we have not really gone much further than this acknowledgement that there is a need to think about the conceptual and discursive frames that accompany these data. It is this project that needs to be attended to, with some urgency. This now needs sustained attention to build upon some of these insights and to reinvigorate the type of project that Foucault, if you will pardon the assumption, may have taken on were he to have been around to observe the emergence or birth of Big Data.

If we were to explore the history of Big Data in terms of the history of the concept then that would lead us to try to understand the work that this concept does to shape practices and behaviours, to limit jurisdictions and to establish truths and desired outcomes. In short, it is to explore the world views or perspectives that the term Big Data is woven from and provokes. This would be to approach the term Big Data as being built in the tensions of veridiction, and to see how it authorises certain behaviours, actions and outcomes. To see what perspectives and notions of truth that it endorses. To see how it brings with it a set of preferences and desires that it then legitimises. This is to see how the term Big Data itself has political ends as it comes to demarcate value or worth. We can certainly start by thinking about how the concept evokes certain feelings of trust through its apparent properties of objectivity and neutrality. In short, *the concept of Big Data frames and makes-up the data themselves*. With this in mind we need to see what type of work it does, how it leads us to see those data and how this framing is woven with particular ways of seeing that social world. The framing of the data is particularly powerful in this regard and will dictate not only what we get from the data but also the possibilities that are afforded simply from uttering these two words together.

This approach will require us to look across different sectors to see how Big Data, the term, has been used. It will look at how it has been deployed in commercial, political, economic and organisational discourse, and what type of work it has done in these sectors. Focusing upon this will hopefully then open up broader political motifs as they find their way into the language of everyday governance and social ordering.

Based on these discussions, Table 1 attempts to summarise the key analytical points that will be required to extend this project. Table 1 provides an analytical framework for exploring the work that is being done by the concept of Big Data. The left hand column presents the analytical focal points and the right hand column presents the types of questions and issues raised by those focal points. These are intended as

Table 1. An analytic framework for structuring the analysis of the work done by the concept 'Big Data'.

Big Data's 'grids of perception'focal points for analysing the making-up and framing of Big Data	Analytical questions and issues
Promises	What promises are being made in the discussion of the data? What hopes and futures are evoked or imagined? How is Big Data seen to promise possible outcomes, efficiencies, improvements and forms of progress?
Manifestations of truth	How is the data to be used to distinguish truth from falsity? What possibilities are presented in the truths implicit in the discussion of the data? What are the truths that Big Data is perceived to enable us to reveal, discover or uncover?
Jurisdiction formation and maintenance	How is Big Data used to set the territories of knowledge? How are boundaries placed around what can be known? Who is responsible for deciding what can be known through the data? How is this position policed and controlled? Who decides what is knowledge and who has the right to use and know it?
Veridiction	How is Big Data seen to present opportunities to verify, authorise and render appropriate? How is Big Data seen to legitimise and justify? What is then seen to be afforded by these legitimising processes? What do these systems verify and why? How does this link to truth making?
The demarcation of value and worth	How is Big Data used to frame what is seen to be valuable or worthwhile? How is Big Data used in the boundary work required to demarcate value? How is Big Data used to promote certain forms of value and to devalue other things? What are the implicit values laced into discussions of Big Data? Can worth be measured?
Limits placed on practice and behaviour	How is Big Data used to justify and present preferred practices and behaviours? How are the limits drawn around the acceptability of behaviours? How is Big Data used to define normality and abnormality? How is the sharing of practices implicated with certain limits that are woven into the discussion of Big Data?
Objectivity and neutrality	How is Big Data presented as being an objective form of knowledge? To what extremes is this form of objectivity taken? How is the data presented as providing the basis of neutral forms of decision making or decision making at a distance? What are the questions of agency raised by the concept of Big Data and how is the responsibility for decision making shifted to these data?
Judgments and evaluations	What types of judgments and evaluations is Big Data seen to make possible? How is Big Data seen to afford processes of judgment and evaluation? What judgments and evaluations are being made and with what types of temporality, frequency and strength of outcome? What opportunities are presented for challenging those judgments? Or is Big Data seen to present the means for incontestable judgment and evaluation?

analytical points of departure that will reveal the implicit dynamics of the concept of Big Data and will thus let us analyse it as a programme of activity and a way of thinking that becomes realised in the limits and practices of the social world. This framework is a heuristic that can be used to guide and shape our analysis, but it may well need to be adapted. The suggestion then is that the framework offered in Table 1 may be used to explore how the concept of Big Data is enacted and performed in making-up data across different social spheres and sectors. In short, this is an analytical framework that may be drawn upon wherever talk of Big Data may be found.

Conclusion

To conclude, I'd like to suggest that we have a relatively pronounced understanding of how data 'make people up', to use Hacking's (1990: 3) term, but we have relatively little appreciation of how *concepts make-up those* *data*. This is not to say that the particular material properties of the current data moment are unimportant – they clearly need further work to understand how those properties relate or differentiate them from that broader history of social statistics. It is to say though that what is also needed is a detailed exploration of the trajectory and influence of the concept of Big Data. We need to ask what work this term does, and what work it has done. We need to explore how it has become established within the discourse of organisations, funding bodies, political and policy circles, in journalism, in social commentary, and in various others sectors. We need to look at the emergence of this powerful concept and to understand how it has been shaped and reshaped in its use. We also need to understand how the term Big Data brings data to life, how it breathes life into data, how it makes them vital and telling.

Underpinning this approach is the pursuit of a more detailed understanding of how Big Data, as a concept, recrafts notions of value and worth. The concept of Big Data might seem unimportant – it might be dismissed as 'business' or 'managerial' talk, it might be seen as a passing fad, it might be seen to be part of the meaningless verbiage of contemporary media cultures – but the scale of the use of the term would suggest something different. The term Big Data is doing a lot of work, it is a persuasive presence in funding, management, decision making, 'human capital' and the everyday practices of production and consumption. The work that is being done by the concept of Big Data needs attention, particularly as it is frequently doing far more than the actual data itself. Indeed, the term Big Data can be used to reveal the type of thinking and the mode of reasoning that ushers data and metric-led processes into everyday, organisational and social life. A part of the role that this concept plays concerns the different ways that it demarcates what is valued or what is seen to be worthwhile. It is a term that lends confidence, authority and objectivity to decisions that are then realised through the data themselves. This gives this particular term a very powerful social presence that needs to be unpicked. The threads then need to be followed back through the history of its usage.

All of this will require us to understand the visions within which and through which notions of Big Data are communicated. To see the way that Big Data is evoked and the kind of outcomes and sensibilities that it provokes. The power of Big Data is not just in the data themselves, it is in how those data and their potential is imagined and envisioned. To understand the power, influence and reach of Big Data requires us to understand the performative influence of the material data whilst also being attentive to the concept that frames them. My suggestion is that so far we have focused virtually all of our attention on the phenomenon and we have given very little attention to the powerful concept that defines, enacts and ushers in those apparently Big Data.

By looking back at some important historical accounts we can quickly see that what is most novel about Big Data is not necessarily the vast accumulation of data, although that is an important part or moment of an established and long set of genealogical threads, but the way that this concept of Big Data has taken on such commercial, organisational and economic force and power. For this reason, amongst others, I would suggest that we now need to lavish some attention on this loaded and powerful concept, particularly as it comes to define contemporary life in so many ways. It has been argued that when thinking of how our lives are measured we need to think about the modes or styles of thought that accompany that measuring, rather than just focusing upon the technical infrastructures (see Elden, 2006: 139–148; Hacking, 1990; Porter, 1986, 1995). This is certainly true for Big Data. The pursuit of Big Data, like the pursuit of statistical measures of populations, is as much about a mode of reasoning or a way of thinking as it is about the assemblage that it generates.

The way to explore these modes of thought or styles of reasoning is to unpick and illuminate the role played by that very label of Big Data in various social spheres. Of course, any clear separation of the material phenomenon from the concept of Big Data is misleading, they work together and are intimately intertwined. My point here is that we need to think about the historical context in which Big Data is unfolding, we need to see it as part of the long series of developments in the measurement of people and populations. At the same time though, in pursuing a more contextual understanding we should not continue to be preoccupied with the data itself we also need to examine the type of data-thinking that is encapsulated in the term Big Data and in the use of that term. There is something to be said for this particular moment in the long unfolding of metric based approaches to the social world, there are likely to be a number of things that are materially distinct about this particular moment in that history, but the things that need to be said require us to understand these apparently new forms of data whilst also paying careful attention to the way that they are packed, presented and rolled-out in the discourse that surrounds and permeates them. This may or may not be a unique or important moment in the history of social statistics and metrics, but it is nevertheless a moment in which a particular concept is taking hold and in which its power is worth some reflection. We simply cannot understand Big Data in historically informed and critical terms unless we analyse the interconnections between its materiality and the concept through which these material transformations are understood. What is perhaps most interesting about this moment in this long history is that we have such a prevalent and prominent term that presents this phenomenon to us as if it was a sudden and unique moment within that history.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. In this article I will not offer a direct definition of 'Big Data' as such. This is for two reasons. First, this type of definition has been provided elsewhere, such as in Rob Kitchin's (2014) excellent and authoritative overview of Big Data (which includes a chapter detailing the definition of this term, see Kitchin, 2014: 67–79). And, second, the approach that I outline in this article aims to explore the meanings and rationalities associated with the term 'Big Data'. As such, it aims to explore the various definitions that are attached to this particular term rather than treating it as a fixed entity. This article actually aims to use the term Big Data as a way into these types of defining statements and understandings – meaning that tying down the meaning too tightly from the outset may hamper its progress and scope.

References

Ajana B (2013) *Governing Through Biometrics: The Biopolitics of Identity*. Basingstoke: Palgrave Macmillan.

Beer D (2016) *Metric Power*. London: Palgrave Macmillan.

Bhambra GK (2014) *Connected Sociologies*. London: Bloomsbury.

boyd D and Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679.

Day R (2014) *Indexing it All: The Subject in the Age of Documentation, Information, and Data*. Cambridge, MA: MIT Press.

Desrosières A (1998) *The Politics of Numbers: A History of Statistical Reasoning*. Cambridge, MA: Harvard University Press.

Elden S (2006) *Speaking Against Number: Heidegger, Language and the Politics of Calculation*. Edinburgh: Edinburgh University Press.

Elden S (2007) Governmentality, calculation, territory. *Environment and Planning D: Society and Space* 25(3): 562–580.

Elden S (2013a) How should we do the history of territory? *Territory, Politics, Governance* 1(1): 5–20.

Elden S (2013b) *The Birth of Territory*. Chicago: Chicago University Press.

Espeland WN and Sauder M (2007) Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology* 113(1): 1–40.

Espeland WN and Stevens ML (2008) A sociology of quantification. *European Journal of Sociology* 49(3): 401–436.

Featherstone M (2000) Archiving cultures. *British Journal of Sociology* 51(1): 168–184.

Foucault M (1991) Questions of method. In: Burchill G, Gordon C and Miller P (eds) *The Foucault Effect*. Chicago: The University of Chicago Press, pp. 73–86.

Foucault M (2007) *Security, Territory, Population: Lectures at the Collège de France 1977–1978*. Basingstoke: Palgrave Macmillan.

Foucault M (2008) *The Birth of Biopolitics: Lectures at the Collège de France 1978–1979*. Basingstoke: Palgrave Macmillan.

Foucault M (2014) *On the Government of the Living: Lectures at the Collège de France 1979–1980*. Basingstoke: Palgrave Macmillan.

Hacking I (1990) *The Taming of Chance*. Cambridge: Cambridge University Press.

Hacking I (1991) How should we do the history of statistics?. In: Burchill G, Gordon C and Miller P (eds) *The Foucault Effect* Chicago: The University of Chicago Press, pp. 181–195.

Halpern O (2014) *Beautiful Data: A History of Vision and Reason since 1945*. Durham, NC: Duke University Press.

Kitchin R (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures & their Consequences*. London: Sage.

Kitchin R and McArdle G (2016) What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society* 3: 1–10.

MacKenzie D (1981) *Statistics in Britain: The Social Construction of Scientific Knowledge*. Edinburgh: Edinburgh University Press.

Manning P (2013) *Big Data in History*. Basingstoke: Palgrave Macmillan.

Porter TM (1986) *The Rise of Statistical Thinking 1820–1900*. Princeton, NJ: Princeton University Press.

Porter TM (1995) *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.

Rottenburg R, Merry SE, Park SJ, et al. (eds) (2015) *The World of Indicators: The Making of Governmental Knowledge Through Quantification*. Cambridge: Cambridge University Press.

Savage M (2013) The 'social life of methods': A critical introduction. *Theory, Culture & Society* 30(4): 3–21.

Schinkel W (2013) The imagination of 'society' in measurements of immigrant integration. *Ethnic and Racial Studies* 36(7): 1142–1161.

Tyler I (2015) Classificatory struggles: Class, culture and inequality in neoliberal times. *The Sociological Review* 63(2): 493–511.

Commentary

Big Data & Society
July–December 2014: 1–5
© The Author(s) 2014
DOI: 10.1177/2053951714539277
bds.sagepub.com

Big Data from the bottom up

Nick Couldry and Alison Powell

Abstract

This short article argues that an adequate response to the implications for governance raised by 'Big Data' requires much more attention to agency and reflexivity than theories of 'algorithmic power' have so far allowed. It develops this through two contrasting examples: the sociological study of social actors used of analytics to meet their own social ends (for example, by community organisations) and the study of actors' attempts to build an economy of information more open to civic intervention than the existing one (for example, in the environmental sphere). The article concludes with a consideration of the broader norms that might contextualise these empirical studies, and proposes that they can be understood in terms of the notion of voice, although the practical implementation of voice as a norm means that voice must sometimes be considered via the notion of transparency.

Keywords

Agency, reflexivity, analytics, political economy, voice, transparency

Introduction

We are living through a transformation of governance – both its mechanisms and reference-points – which is likely to have profound implications for practical processes of government and everyday understandings of the social world. A shift is under way from discrete forms of intervention in social space based on intermittent and/or specific information-gathering to continuous processes of management based on total and unremitting surveillance (Ruppert, 2011). Both management and government increasingly are becoming predicated upon the continuous gathering and analysis of dynamically collected, individual-level data about what people are, do and say ('Big Data'). However misleading or mythical some narratives around Big Data (Boyd and Crawford, 2011; Couldry, 2013), the actual processes of data-gathering, data-processing and organisational adjustment associated with such narratives are not mythical; they constitute an important, if highly contested, 'fact' with which all social actors must deal. This article will offer a social approach to the construction and use of such data and related analytics.

The possibility of such a social approach to Big Data has, until now, been obscured by unnecessarily generalised readings of the consequences of these broad changes. Without a doubt, the information types that management and governance take as their starting-point have changed: it is digital infrastructures of collection, transmission, analysis and presentation that have made possible continuous data-mining. Compared to representative sampling, such new approaches to data collection are totalising; they are also characterised by the aggregation of multiple data sets through the use of calculation algorithms. This seemingly increased role for algorithms has led some commentators to focus on the dominance of 'algorithmic power' (Lash, 2007), an approach that leaves no room for agency or reflexivity on the part of 'smaller' actors. We posit that emerging cultures of data collection deserve to be examined in a way that foregrounds the agency and reflexivity of individual actors as well as the variable ways in which power and participation are constructed and enacted.

London School of Economics, London, UK

Corresponding author:
Nick Couldry, Department of Media and Communications, London School of Economics, Houghton St., London WC2A 2AE, UK.
Email: N.Couldry@lse.ac.uk

This more agent-focused inquiry into the consequences of algorithmic calculation's deep embedding in everyday life has been foreshadowed in some earlier debates, notably Beer's (2009) response to Lash's (2007) argument that 'algorithmic power' has changed the nature of hegemony. As Beer (2009: 999) noted, sociology must also 'focus . . . on those who engage with the software in their everyday lives'. Such a focus does not come naturally within Lash's broadly philosophical formulations of issues in social theory which foreground 'a collapse of ontology and epistemology' (Lash, 2006: 581), and a new power-laden regime of 'facticity' (Lash, 2007: 56) in which 'there is no time, nor space . . . for reflection' (Lash, 2002: 18). If that were right, why pay close attention to what actors say when they 'reflect' on their position in the social world? But this analytic closure is unhelpful.

Needed instead is a more open enquiry into what actual social actors, and groups of actors, are doing under these conditions in a variety of places and settings. Without denying of course the 'generative' importance of algorithms (Lash, 2007: 71) when embedded in modes of calculation, processing and rule, we need to remember that social actors are often themselves aware of being classified. Even if they are not privy to the details of when, by whom, and how they have been classified, that this has happened is something of which they are aware, and indeed one of the main 'facts' they have to deal with as social actors. We need to become sensitive to what Beer (2009: 998) has called people's 'classificatory imagination' and, over the longer term, the wider 'social imaginaries' (Mansell, 2012; Taylor, 2005) that may be emerging around these new cultures of data collection.

Beer goes on helpfully to distinguish three levels of resulting empirical research: first, regarding the 'organizations that establish and activate Web 2.0 applications'; second, regarding the 'software infrastructures and their applications on the web'; and third, regarding how the first two levels 'play out in the lives of those that use (or do not use) particular web applications' (2009: 998). We would like in this short article to build particularly on Beer's third level, and on the lessons of our own empirical researches, to map out some more detailed and concrete ways of researching the everyday uses of data and analytics from a social perspective. The result is to open up a much wider and more varied space of agency and reflexivity than allowed for in philosophical accounts. The likely outcome may be no less critical of Big Data's implications, but will develop critique through a more nuanced characterisation of 'Big Data' as a variegated space of action, albeit one very different from the spaces in which pre-digital social actors operated.

Doing social analytics

Our first example of a more agent-focused account of Big Data is what has been called 'social analytics' (see Couldry et al., forthcoming, for a much more detailed account). A social analytics approach is an explicitly sociological treatment of how analytics get used by a range of social actors. Such an approach aims to capture how particular actors reflect upon, and adjust, their online presence and the actions that feed into it, through the use of 'analytics'. 'Analytics' here is used broadly to cover both basic analytics (the automated measurement and counting installed within the operation of digital platforms and associated websites, apps and tools) and the adjustments made by actors themselves in response to such measurement and counting operations. Platforms that count and sort online data, such as Google and Facebook, work automatically via algorithms, often allowing users only limited degrees of manual adjustment (van Dijck, 2013). Other adjustments around those operations may take direct digital form (a website redesign) or organisational form (an adjustment in an organisation's management of its resources). In all these cases, the variable use of analytics is a social process involving reflection, monitoring and adjustment.

By 'social actors' we mean actors with social ends over and above the basic aim of generating and analysing data (usually for profit): that basic aim in itself is of little sociological interest. The broader sociological interest starts when there is some tension, actual or potential, between the aims that social actors are trying to achieve and the interpretations of their activities that analytics generate. This use of the term 'social analytics' encompasses, but goes beyond, the everyday 'technical' use of the term 'analytics' to mean the measurement and reporting of internet data. The mutual intertwining of human and material agency is hardly a new insight (Pickering, 1995: 15–20), but it acquires a special interest when analytics' operations are opaque to non-expert social actors who must work hard to acquire control over them.

One key variable in such research is what is measured and analysed, the 'object' of analytics. The underlying data's relationship to an organisation's online presence may be more or less direct: direct if the data is literally about that organisation's online presence (numbers of unique users, their characteristics, types of interaction with online content); or indirect if the data is not about an organisation's online presence, but is generated or presented online, becoming part of how that organisation is judged by online visitors (online reviews, debates). The closeness, or distance, of the relation between the object of data analysis and the general aims and practice of social actors clearly will shape the degree of tension and reflexivity that exists over the implementation of analytics. At one end of the spectrum will be cases where analytics are used directly to support other mechanisms of power (e.g. performance management); at the other end will be cases where what is at stake in the use of analytics is the broad redefinition of an organisation's aims and performance, with no direct impact on the evaluation or management of individuals. In the former case, social analytics may merge into the study of management and power; in the latter case, social analytics may be something closer to a phenomenology of how social actors and organisations with social aims appear to themselves, and to the world, under digital conditions.

Other variables when doing social analytics will include the degree of technical expertise of the actors involved, including the degree to which they can go beyond merely using off-the-shelf analytics to customising them, or perhaps even developing their own analytic tools and data-collection designs. Financial and other resources will also affect how far the processes which social analytics studies can develop, or get blocked, for example, if the staff to do the analytic work that would enable a richer re-evaluation of an organisation's digital presence cease to be available. Expertise and resources are, of course, variables in any fieldwork setting.

Within these basic parameters, however, social analytics promise a rich vein of inquiry into the conditions of data use and analytics use, from the perspective of social actors who are not principally experts in relation to data or algorithms, but who look to them to do certain work towards other ends. It has so far been explored in the context of community and civic activism, but it has the potential to be expanded to many more areas.

Data as media

For media scholars more generally, the shift to a data-rich environment poses challenges for a robust understanding of how agency and expression might still work within that environment. The critical tradition in media and communications has largely been concerned with the operation of power in the construction of systems of symbolic mediation – for example, the function of ideological systems (in the Marxist tradition) or the Gramscian concept of hegemony. These strategies have allowed media and communication scholars to 'work backwards' through systems of symbolic mediation in order to understand the process and initial starting points of mediated 'messages'. This focus on the symbolic quality of media messages allows us to examine power relationships from several different vantage points. Within traditional broadcast media forms we can observe how the symbolic control of mediated messages solidifies control and results in things like propaganda, but we can also see how alternative media producers can wrest control of ideas and their representation to challenge that kind of hegemony.

Broadcast models have however been overtaken, for important purposes, by models of mass self-communication. Whereas institutionalised mass media is structured to disseminate messages from one to many, mass self-communication is structured to invite continual input of data by individuals. This reorganisation of media production initially seemed to promise a reconfiguration of the top-down production of ideology and the bottom-up resistance to it, but as political–economic analyses have developed, we are beginning to see how such shifts have also led to the production of data replacing the production of audiences.

If the exemplary product of institutionalised mass media is propaganda, the exemplary product of mass self-communication is data. A mass media apparatus requests information to be disseminated from the one to the many; its economic model uses this information to generate an audience whose attention can be sold to an advertiser. In the mass self-communication model individuals are still part of an aggregate product to be sold, but instead of their attention on a single message produced for broadcast, it is their individual acts of communication that comprise the 'Big Data' and drive much media value-extraction.

Early critics of mass self-communication noted that the model encouraged individuals to create 'content' that was then sold to others in order to capture their attention (Terranova, 2000; van Dijck, 2013). However, 'content' is still expressive, even when it is sold to capture attention. A more complicated issue concerns the data that is produced, often unwittingly, which now generates much of the value in the newest iteration of the contribution economy. Many everyday activities now produce data without requiring human meaning-construction (or even basic consent). The rise of sensor networks has meant that increasingly individuals are producing not 'content' composed of messages containing intrinsic or constructed meaning, but mere data – temperature readings, status updates, location coordinates, tracks, traces and check-ins. Not one of these individual data-types is necessarily meaningful in itself – but taken together, either through aggregation, correlation or calculation, such data provide large amounts of information. The difference between this and the 'content' that mass self-communication promises to distribute is that the meaning of data is made not semantically (through expression and interpretation) but through processing – especially the matching of metadata (Boellstorf, 2013). Big Data sets are composed of numerous pieces of information that can be cross-compared, aggregated and disaggregated and made very finely grained, not things whose creators necessarily endowed with meaning. In mining the data, more insights are made available about more aspects of everyday life but no opportunity is provided for these insights to be folded back into the experience of everyday life. In this context, is there any scope, as Boellstorf urges, for integrating the epistemic perspectives of ethnography back into the calculative logic of meta-data?

All along, the political economy of personal data, as anticipated by Gandy (1993), has been concerned with value created through the aggregation and calculation of individual traces. Even if we leave aside the expressive quality of individual acts of communication online, the production of data as a by-product of everyday life practices enacts a particular political economics of media, undertaken within a situation of pervasive surveillance and generalised authoritarianism (Cohen 2012). But the potential disconnect between system and experience, phenomenology and political economy, can be overcome by examining on the ground agents' strategies for building alternative economies of information. Such alternative economies are being developed in several areas related to environment and sustainability, including projects that use data sources to make provenance and supply chains visible, and those that encourage individuals and communities to collect data as a means to make environmental issues visible by challenging conventional data collection.

Academic projects like Wikichains (Graham, 2010) and start-up companies like Provenance.it (2013) aggregate various forms of data about the production, distribution and supply chains of manufactured objects as a means of drawing attention to their long-term ecological and economic costs. While Provenance.it remains anchored in a consumer-based economic model, it does illustrate how alternative modes of data collection and analysis could shift agency and representation, especially if it permitted for greater reflexivity. Similarly, NGOs like Mapping for Change (2013) have supported individuals and community groups in gathering environmental data (like air quality and noise) as a means of engaging with gaps and flaws in official data. These actions intervene in efforts to use such environmental data within top-down governance processes. As Gabrys (2014) identifies, such citizen science efforts must be enfolded and imagined in processes of environmental governance or 'biopolitics 2.0'. These examples illustrate two ways that an alternative economics of information might employ calculation of multiple data sources or generation of alternative sources to illustrate or critique power relations, although they also illustrate the ambiguity of accountability within these processes.

Voice, transparency and power

The rise of analytics presents a significant normative challenge for scholars, activists and others who seek to understand how humanity, sociability and experience are represented. The daily practices of grappling with data and with the consequences of data analyses generate new questions about what and whose power gets exercised through such practices, and to what degree such exercises of power are satisfactorily made accountable. One approach to these challenges is through attention to problems of voice (Couldry, 2010). Voice, understood as a value for social organisation (Couldry, 2010: ch. 1), involves taking into account agents' practices of giving an account of themselves and their conditions of life. The value of voice is essential to the workings of any models so far developed of democratic institutions, but it is not immediately compatible with a world saturated with the automated aggregation of analytic mechanisms that are not, even in principle, open to any continuous human interpretation or review.

While the notion of voice insists upon organisational processes being accountable to the subjectivities and expressiveness of all, the movement towards more casual, automatic sensing and its calculative rather than epistemic logic seems to eliminate this accountability. Yet clearly something similar to 'voice' is required in this new world, and this is not just a matter of democracy: 'we have no idea', wrote Paul Ricoeur, 'what a culture would be where no one any longer knew what it meant to narrate things' (Couldry, 2010: 1, quoting Ricoeur, 1984: 29). At present, the proxy for voice in the algorithmic domain is the notion that data gathering processes ought to be transparent, and the logic of calculation revealed. A focus on transparency could begin to foreground notions of accountability in data calculation, ownership and use.

Notions of transparency have been discussed with respect to government production and use of data (Tkacz, 2012). Yet despite pledging to make public data collection transparent, governments like the US and the UK in fact collect much more information via surveillance projects and partnerships with information technology companies. With the reform of the USA's National Security Administration, perhaps more attention will begin to be paid to the data collection practices of the technology sector, making more of them visible. This kind of transparency goes part of the way to establishing accountability, but it still fails to address accountability and reflexivity. A refined concept of transparency that is sensitive to the meaning that data trails might form (even if it cannot be sensitive to the meaning inherent in their production) might go some way to addressing this. This is a tricky proposal: unless and until the unconscious production of data can be conceived of as a form of expression, the philosophical basis for such an expansive transparency will be difficult to establish. One possible way to proceed might be to highlight not just the risks of creating and sharing data but the opportunities as well. The practices of social analytics and citizen science have the potential to establish these opportunities, ambiguous as they may be.

We hope that, as the debates about Big Data and society continue and their democratic stakes become clearer, the values implicit in the terms ‘voice’ and ‘transparency’ will themselves begin to converge in more satisfying ways than are at present possible.

Declaration of conflicting interest

The authors declare that there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Beer D (2009) Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society* 11: 985–1002.

Boellstorff T (2013) *Making Big Data, in Theory*. First Monday, [S.l.], September 2013. ISSN 13960466. Available at: http://firstmonday.org/ojs/index.php/fm/article/view/4869/3750 (accessed 27 January 2014).

boyd d and Crawford K (2011) Critical questions for Big Data: Provocations for a cultural, technological and scholarly phenomenon. *Information, Communication and Society* 15(5): 662–679.

Castells M (2009) *Communication Power*. Oxford: Oxford University Press.

Cohen J (2012) *Configuring the Networked Self*. New Haven: Yale University Press.

Couldry N (2010) *Why Voice Matters*. London: Sage.

Couldry N (2013) *A Necessary Disenchantment: Myth, Agency and Injustice in a Digital Age*. Inaugural lecture at LSE. Available at: http://www.lse.ac.uk/newsAndMedia/videoAndAudio/channels/publicLecturesAndEvents/player.aspx?id=2120 (accessed 21 November 2013).

Couldry N, Fotopoulou A and Dickens L (forthcoming). *Real Social Analytics: A Contribution Towards the Phenomenology of a Digital World*.

Gabrys J (2014) Programming environments: Environmentality and citizen sensing in the smart city. *Environment and Planning D: Society and Space* 32(1): 30–48.

Gandy O (1993) Toward a political economy of personal information. *Critical Studies in Mass Communication* 10(1): 70–97.

Graham M (2010) *‘WikiChains: Encouraging Transparency in Commodity Chains’ Research Project*. Available at: http://www.oii.ox.ac.uk/research/projects/?id=75 (accessed 30 May 2014).

Lash S (2002) *Critique of Information*. London: Sage.

Lash S (2006) Dialectic of information? A response to Taylor. *Information Community & Society* 9(5): 572–581.

Lash S (2007) Power after hegemony: Cultural studies in mutation. *Theory, Culture & Society* 24(3): 55–78.

Mansell R (2012) *Imagining the Internet*. Oxford: Oxford University Press.

Mapping for Change (2013) *Services: Citizen Science*. Available at: http://www.mappingforchange.org.uk/services/citizen-science/ (accessed 30 May 2014).

Pickering A (1995) *The Mangle of Practice*. Chicago: Chicago University Press.

Provenance.it (2013) *About Provenance*. Available at: https://www.provenance.it/about (accessed 30 May 2014).

Ricoeur P (1984) *Time and Narrative*, Vol. 2. Chicago: Chicago University Press.

Ruppert E (2011) Population objects: Interpassive subjects. *Sociology* 45(2): 218–233.

Taylor C (2005) *Modern Social Imaginaries*. Durham, NC: Duke University Press.

Terranova T (2000) Free labor: Producing culture for the digital economy. *Social Text* 18(2): 33–58.

Tkacz N (2012) From open source to open government: A critique of open politics. *Ephemera* 12(4). Available at: http://www.ephemerajournal.org/contribution/open-source-open-government-critique-open-politics-0 (accessed 30 May 2014).

Van Dijck J (2013) *The Culture of Connectivity*. Oxford: Oxford University Press.

16 COMMENTS

Craig Dalton and Jim Thatcher – What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'

What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data.'

*Co-authored by **Craig Dalton**, Assistant Professor of Geography, University of Bloomsburg and **Jim Thatcher**, Assistant Professor, University of Washington-Tacoma (listed alphabetically)*

(https://societyandspace.files.wordpress.com/2014/05/critical_data_studies.jpg)There is a need for a critical data studies

"The benefits to society will be myriad, as big data becomes part of the solution to pressing global problems like addressing climate change, eradicating disease, and fostering good governance and economic development." (Mayer-Schoenberger and Cukier, 2013: 17)

"A statistical model of society that ignores issues of class, that takes patterns of influence as givens rather than as historical contingencies, will tend to perpetuate existing social structures and dynamics. It will encourage us to optimize the status quo rather than challenge it." (Carr, 2014)

As the public discourse around data turns from hubristic claims to existing, empirical results, it's become nearly as easy to bash 'big data' as to hype it (Carr, 2014; Marcus and Davis, 2014; Harford, 2014; Podesta, 2014). Geographers are intimately involved with this recent rise of data. Most digital information now contains some spatial component (Hahmann and Burghardt, 2013) and geographers are contributing tools (Haklay and Weber, 2008), maps (Zook and Poorthius, 2014), and methods (Tsou et al. 2014) to the rising tide of quantification. Critiques of 'big data' thus far offer keen insight and acerbic wit, but remain piecemeal and disconnected. 'Big data's' successes or failures as a *tool* are judged (K.N.C. 2014), or it is examined from a specific perspective, such as its role in surveillance (Crampton et al. 2014). Recently, voices in critical geography have raised the call for a systemic approach to data criticisms, a critical data studies (Dalton and Thatcher, 2014; Graham, 2014; Kitchin, 2014). This post presents seven key provocations we see as drivers of a comprehensive critique of the new regimes of data, 'big' or not. We focus on why a critical approach is needed, what it may offer, and some idea of what it could look like.

1. **Situating 'big data' in time and space**

Data has always been big. Such a phrase borders on the trite, but it is important to recognize the epiphenomenal nature of the term 'big data.' It is specific to a moment in time whose dominance seems already to be shifting. This does not mean that 'big data' is going away. Much as the term "e-commerce" disappeared from our conscious use as online shopping became a normal practice (Leyshon et al. 2005), 'big data' is simply receding into the banality of the every-day. This enables and constrains sets of social processes, but does so from an unconsidered position that rises to our attention only when it fails (Harman, 2010). In doing so, 'big data' appears inevitable, naturalizing its consequences and foreclosing alternative possibilities. To understand 'big data' and whatever comes next, we must resist this urge to let it stand apart from history and pass silently into our everyday lives.

'Big data' has big precursors, earlier knowledges that set the stage and helped define the nature and needs that present-day 'big data' realizes. The epistemologies of Nineteenth Century statistical mapping (Schulten, 2012), social physics and geography's quantitative revolution (Barnes and Wilson, Forthcoming), the development of geodemographic targeted marketing (Goss, 1994), and the boom-bust cycle of the information technology industry all laid the conditions that realized 'big data.' Today, as 'big data' is enrolled in social processes, it also facilitates power geometries between companies – such as Google, Acxiom, and Foursquare – agencies – such as the NSA – and consumer citizens. We must ask: Whose data? On what terms? To what ends? Attempts to set aside or ignore 'big data's' ancestry and effects serve to hype it, but not better understand it. Situating 'big data' knowledges help us understand both what is happening and why.

2. **Technology is never as neutral as it appears**

As the pushback against 'big data' begins, its excremental qualities (Pearce, 2013) focus around its limitations: the reality of what technology can do versus grandiose claims and hype. In these critiques, 'big data' is a tool and its failures are found in its inability to perform its supposed function – to model and predict reality along certain positivist lines (Harford, 2014). By doing so, these critiques fall within the same epistemological frame of 'big data' itself.

'Big data,' as a technology, is never a neutral tool. It always shapes and is shaped by a contested cultural landscape in both creation and interpretation. Whether in critique or celebration, an instrumental examination of 'big data' will necessarily miss its underlying epistemological effects. The myths of 'big data' are myths that suffuse modern society, seeping into ideas of the quantified self and smart cities. As the fullness of human experience in the world is reduced to a sequence of bytes, we should not limit our concern to how much better those bytes function vis-à-vis their counterparts. Rather, we must ask what it means to be quantified in such a manner, what possible experiences have been opened and which have been closed off? How is 'big data' as a form of technology enabling and constraining our culture and our lives?

Citing Tony Benn, a British Labour party politician, Mark Graham recently suggested we ask of 'big data' "What power have you got? Where did you get it from? In whose interests do you exercise it? To whom are you accountable? And how can we get rid of you?" (Graham, 2014). Just as the so-called "science wars" taught us to question the processes by which austere scientific knowledge is produced, we must also question 'big data.' Quantified digital information, whether called 'big data' or not, is here to stay. As with all successful technologies, it recedes from our attention as it saturates and structures our everyday lives (Feenberg, 1999; Kitchin and Dodge, 2011). We must critically ask who it speaks for and why before it disappears from consideration. To do so, we "follow the [data] scientists" (Latour, 1988).

3. **'Big data' does not determine social forms: confronting hard technological determinism**

Technological change and society have an intricate, recursive relationship. 'Big data' and its concept of data has a role in today's social changes, but it is more complex than simple consequences of large, fast, individualized data analytics or attempts to model society. The innovation, production, and popular use of a technology occurs within and reflects a social context shot through with power, economies, identities, and biases. Even as technology and buzzwords change rapidly, the wider societal processes that shape technology and give it purpose show only gradual change. The popularization of 'big data,' the hype around it, and the backlash against it owe much to the pre-existing needs of ever-growing capital accumulation and crises of legitimacy among public agencies.

A technology does not act alone, out of context, determining the form of society. It plays an ensemble role in social changes as it is utilized for one social purpose or another, facilitating material changes in the structure of society and peoples' everyday lives and deaths. As something made by and for people, a new technology is designed to fulfill social imperatives, such as accumulating capital. In practice, technology can be deployed by many different kinds of people, opening new possibilities (Haraway, 1991) and networks (Terranova, 2004).

A technology designed by one group of stakeholders for a particular purpose may be adopted by different stakeholders and used against its original intended function. In some cases, stakeholders may even reject a technology or pass it by in favor of something else. These political projects and resistances enable and constrain the social and material possibilities down the line (Feenberg, 1999; 2002). Some consumer subjects already attempt to resist aspects of 'big data' using pseudonyms, private web browsing, ad/script blocking, location spoofing, web proxies or VPN services, and turning off location services on their mobile devices. 'Big data's' incomplete, contested nature marks it as much the product of society as society's producer.

4. **Data is never raw**

'Big data' is the result of a specific technological imaginary that rests on a mythological belief in the value of quantification and faith in its ability to model reality (Boyd and Crawford, 2012). In this imaginary, life can be fully captured, quantified, and modeled as theory takes a backseat to 'raw' number crunching. However, in both its production and interpretation, all data – 'big' included – is always the result of contingent and contested social practices that afford and obfuscate specific understandings of the world. The data of 'big data' can take many forms for many purposes: from the massive streams generated by the Large Hadron Collider to the global corpus of tweets. In each case, the data's format and content have been shaped and created for a purpose. Each data model structures and encodes information in one way or another according to the visions of the team of data engineers, scientists, and developers that created it. Furthermore, what is captured is determined by the goals of the project and the analytical model created to instantiate those goals. Fields are defined, accuracies of measurements determined, and other technically necessary steps are taken to create the infrastructure of 'big data.' What is quantified, stored, and sorted? What is discarded? All datasets are necessarily limited representations of the world that must be *imagined as such* to produce the meaning they purport to show (Gitelman, 2013).

Social context is fundamental in both the production and interpretation of meaning. A young boy rapidly contracting his eyelid may be winking, attempting to remove a dust mote, or something else entirely (Geertz, 1973). Ever-present cultural regimes of interpretation structure the analysis of all data, 'big' or small (Boellstorff, 2013). Three different "likes" on a *Facebook* status may reflect three disparate emotional responses: from intense agreement to sardonic recognition to sympathetic pity. However, when it is analyzed simply as a "like" (or an eyelid contraction), the *thickness* of the data and its variety of meanings is lost. In practice, data are not simple evidence of phenomena, they are phenomena in and of themselves (Wilson, 2014) 'Big data' is never "raw." It has always been "baked" through both its construction and its resulting interpretation (Gitelman, 2013). If we are to understand 'big data,' and specifically 'big data' derived from social media, we must engage directly with the cultural regimes of production and interpretation to restore the thick, rich fullness of description that reveals subjects' understandings and intent.

5. **Big isn't everything**

Chris Anderson's (2008) claim that 'big data' meant the "end of theory," where numbers speak for themselves, has become a shibboleth among the 'big data' savvy. Even for data science evangelists like Nate Silver, counterpointing Anderson's hubristic framing of 'big data' serves as a useful way to pivot towards acknowledging the continuing importance of models and theory as "[n]umbers have no way of speaking for themselves" (Silver, quoted in Marcus and Davis, 2013). As the backlash against 'big data' increasingly stresses the importance of domain knowledge, the ability to build sound models from theoretical insights continues to carry weight in practice.

Even with models and theory, 'Big data' analytics cannot answer every research questions, and therefore cannot supplant other, more established qualitative and quantitative research methods. Some propose that researchers can understand the "human dynamics" of a landscape by analyzing 'big data' sets derived from websites, social media and mobile devices (Tsou et al. 2014). "The new [Human Dynamics in the Mobile Age] research agenda may facilitate the transform[ation] of human geography study from qualitative analysis to computational modeling, simulation, and prediction applications using both quantitative and qualitative methods" (Tsou, 2014). Location-tracking someone's phone or tweets may give some trace accounting of their affinity for a place or spatial process, leading to valuable contributions to geographic knowledge (Humphreys et al. 2014). Nevertheless, such a 'big data' approach can never provide the depth and detail that comes with qualitatively learning about and understanding someone's standpoint by actually asking them about a place and their personal feelings and motivations, much less experiencing that place and context for yourself with fieldwork. Purely 'big data' approaches falter with issues of interpretation and context precisely because data is never raw.

A more common charge levelled against 'big data' is that it typically identifies mere correlations in datasets. Further, such large, diverse datasets may be biased. The difference between correlation and causation as well as the care that goes into identifying worthwhile datasets continue to hold validity in an era of 'big data' (Harford, 2014). Likewise, perennial questions of credibility and quality control in geographic data are no less an issue for 'big data' (Goodchild, 2013). Proponents of 'big data' urge us not to rush to judgment as 'big data' analytics continue to develop and may include more robust analyses in the future (Hidalgo, 2014).

Like older quantitative methods that often rely on correlation, such as linear regression, 'big data' analytics are better suited to quantitative questions of *what*, *where*, and *when* than to questions of *how*,and *why*. Analysis of twitter data can map where and when tweets were tweeted and retweeted about a riot following an NCAA basketball championship game, but it cannot answer why individuals chose to tweet or not. In fact, those who did not tweet (or do not ever tweet) remain entirely invisible to the data set. This is neither an unknown (Crampton et al. 2013) nor a paralyzing problem. By comparison, GIS-based quantitative spatial analysis has done profound work with what is a quite limited set of concepts and tools (Pavlovskaya, 2006). More importantly, studies involving GIS also expanded into new and significant areas when they began to include participatory and qualitative approaches (Cope and Elwood, 2009; Craig et al. 2002; Sieber, 2006). Geography is a discipline rich with mixed method approaches, many the result of the joining of empirical and theoretical work made possible when researchers "step[ped] outside of their comfort zones" (Wright et al.1997).

We believe 'big data' research can be similarly improved by working with, rather than denying the importance of, "small data" (Kitchin and Lauriault, 2014; Thatcher and Burns, 2013) and other existing approaches to research. Employing this combined approach requires an awareness among the researchers of the forms of knowledge being produced and their own role in that process. Furthermore, doing critical work with 'big data' involves understanding not only data's formal characteristics, but also the social context of the research amidst shifting technologies and broad social processes. Done right, 'big' and small data utilized in concert opens new possibilities: topics, methods, concepts, and meanings for what can be understood and done through research.

6. **Counter-Data**

What is to be done with 'big data?' Data's role in targeted marketing and the surveillance state are clear, but what other purposes could it serve? The history, discourses, and methods of counter-mapping suggest one opening for critical engagement using 'big data.' Maps have long been a geographic knowledge of imperialism and massive capital accumulation, a means to facilitative exploitative material relationships and proposition our consent to those relationships (Crampton, 2010; Wood, 2010). Much like 'big data,' if maps are judged by these standards alone, hope for critically-informed use appears dim. However, another aspect of mapping is a beautiful diversity of cartographic knowledges that differ from and even run counter to cartography's traditional purposes. Harris and Hazen describe how counter-mapping "challenge[s] predominant power effects of mapping" and "engages in mapping that upset[s] power relations" (2005). Counter-mapping works from the bottom-up within a given situation and includes mapping for indigenous rights (Peluso, 1995), autonomous social movements (Holmes, 2003; Dalton and Mason-Deese, 2012) and art maps (Wood, 2010; Mogel and Bhagat, 2007). In such cases, researchers must be self-conscious of their own positionality and the consequences of knowledge production. Recent work on indigenous mapping makes clear the limits of counter-mapping (Wainwright and Bryan, 2009). Nevertheless, eschewing 'big data' entirely for its ties to surveillance, capital, and other exploitative power geometries forecloses the possibility of liberatory, revolutionary purposes. We must ask what counter-data actions are possible? What counter-data actions are already happening?

7. **What can Geographers do? What is our praxis?**

Approaching 'big data' critically constitutes an opportunity for geographers. Corporations and government agencies include basic spatial criteria into their 'big data' analytics and geographers are already utilizing 'big data' in their research, though predominately in the form of data fumes (Thatcher, 2014). By situating 'big data' technologies and data in contexts and thereby assessing its contingent, non-determinative role and impacts in society, critical data studies offer a less-hyped but more reasoned conceptualization of 'big data.' From this critical standpoint, 'big data' and older 'small data' approaches may be utilized together for better research. Crucially, the critical standpoint also opens possibilities, new questions and topics previously invisible in 'big data' practice. Given this situation, we suggest geography sits at a unique position to help develop a fully critical data studies for three reasons:

First, geographers have decades of experience in analyzing data in terms of space. With the majority of digital information containing a spatial component (Hahmann and Burghardt, 2013), geographic analytical concepts, methods, and models are directly relevant in producing an understanding that data. Furthermore, geographers have also developed critical approaches to spatial analysis, such as Bunge's geographical expeditions (Bunge, 2011; Merrifield, 1995), critical GIS, qualitative GIS, and the above-mentioned counter-mapping. Finally, GIS and cartographic design have prepared geographers for the problems of processing and visualizing complex spatial data for diverse audiences.

Second, geographers emphasize not only space, but *place*. In a world of quantified individualization, understanding the contextual value of place is significant and powerful. Relying solely on 'big data' methods can obscure concepts of place and place-making because places are necessarily situated and partial. Understanding the "making and maintenance of place" remains a central task for geographers (Tuan, 1991: 684), as do the power geometries of *places* and *spaces* (Massey, 1993). Drawing from the traditions of spatial theorists like Tuan, Massey, and Cresswell, geographers are uniquely suited to heed recent calls for more relational understandings of space and place in 'big data' (Crampton et al. 2013).

Third, geography has long been a field that accommodates a broad range of approaches and mixed methods research. For example, studying the connection between natural and social processes is core to the discipline (Yeager and Steiger, 2013). Debates over the nomothetic or idiographic production of knowledge, perhaps most famously found in the Hartshorne-Shaefer debates of the 1950s, have given way to a multitude of methodologies, many informed by both qualitative and quantitative approaches. Critical data studies must build on these hard-learned lessons of theory and praxis. 'Big data's' multidisciplinary nature provides geographers fertile ground upon which to learn from and contribute to other fields like the Digital Humanities and Critical Information Studies (Vaidhyanathan, 2006).

Geography and geographers have much to offer and much to gain from critical data studies, but it is essential to seize the moment before it passes. Much like the advent of Critical Geographic Information Systems, we must engage in the "hard work of theory" (Pickles, 1997). As the term 'big data' normalizes itself in discourse, it recedes from conscious consideration. Now, while 'big data' is still a contested concept in public and academic debates, we must question and challenge its role in an emerging hegemonic order of societal calculation. In this pursuit, we conclude with five questions for critical data studies, some already partially taken up, but all requiring further study:

- What historical conditions lead to the realization of 'big data' such as it is? (Barnes and Wilson, forthcoming; Dalton, 2013)
- Who controls 'big data,' its production and its analysis? What motives and imperatives drive their work? (Thatcher, 2014)
- Who are the subjects of 'big data' and what knowledges are they producing? (Haklay, 2012)
- How is 'big data' *actually* applied in the production of spaces, places and landscapes? (Kitchin and Dodge, 2011)
- What is to be done with 'big data' and what other kinds of knowledges could it help produce? (Shah, 2014)

References

Anderson C 2008 The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired 16*(7).

Barnes T and M Wilson. Forthcoming. Big Data, Social Physics and Spatial Analysis: The Early Years. *Big Data & Society*.

Boellstorff T 2013. Making Big Data, in Theory. *First Monday* 18(10). http://firstmonday.org/ojs/index.php/fm/article/view/4869/3750 (http://firstmonday.org/ojs/index.php/fm/article/view/4869/3750) Accessed May 15, 2014.

Boyd D and K Crawford 2012 Critical Questions for Big Data. *Information, Communication & Society 15*(5): 662-679.

Bunge W 2011(1971) *Fitzgerald: Geography of a Revolution.* Athens, Georgia: University of Georgia Press.

Carr N 2014 The Limits of Social Engineering. *MIT Technology Review.* http://www.technologyreview.com/review/526561/the-limits-of-social-engineering/ (http://www.technologyreview.com/review/526561/the-limits-of-social-engineering/) Accessed May 1, 2014.

Cope M and S Elwood (eds) 2009 Qualitative GIS: A Mixed Methods Approach. Thousand Oaks, CA: SAGE Publications.

Craig W, T Harris, and D Weiner (eds) 2002 *Community Participation and Geographical Information Systems*. CRC Press.

Crampton J 2010 *Mapping: A Critical Introduction to GIS and Cartography*. New York City: Blackwell Publishing.

Crampton J, M Graham, A Poorthuis, T Shelton, M Stephens, M Wilson, and M Zook 2013 Beyond the Geotag: Situating 'Big Data' and Leveraging the Potential of the Geoweb. *Cartography and Geographic Information Science 40*(2): 130-139.

Crampton J, S Roberts and A Poorthuis 2014 The New Political Economy of Geographical Intelligence. *Annals of the Association of American Geographers 104*(1): 196-214.

Dalton C 2013 Sovereigns, Spooks, and Hackers: An Early History of Google Geo Services and Map Mashups. *Cartographica 48*(4): 261-274.

Dalton C and J Thatcher 2014 Inflated Granularity: The Promise of Big Data and the Need for a Critical Data Studies. Presentation at the Annual Meeting of the Association of American Geographers, Tampa, Fl, April 9, 2014. http://meridian.aag.org/callforpapers/program/AbstractDetail.cfm?AbstractID=56048 (http://meridian.aag.org/callforpapers/program/AbstractDetail.cfm?AbstractID=56048) Accessed May 15, 2014.

Dalton C and L Mason-Deese 2012 Counter(Mapping) Actions: Mapping as Militant Research. *ACME 11*(3).

Feenberg A 2002 *Transforming Technology*. New York: Oxford University Press.

Feenerg A 1999 *Questioning Technology*. New York: Routledge.

Geertz C 1973 Thick Description: Toward an Interpretative Theory of Culture. In *The Interpretation of Cultures: Selected Essays* by Clifford Geertz. Basic Books. 3-32.

Gitelman L (ed) 2013 *"Raw Data" Is an Oxymoron*. Cambridge MA: MIT Press.

Goodchild M 2013 The Quality of Big (Geo)data. *Dialogues in Human Geography 3*(3) 280-284.

Goss J 1994 Marketing the New Marketing: The Strategic Discourse of Geodemographic Information Systems. In: John Pickles (ed) *Ground Truth: The Social Implications of Geographic Information Systems*. New York: Guilford Press. 130-170.

Graham M 2014 My Response to the Geoweb and 'Big Data' alt.conference at #AAG2014. *Zero Geography* blog. http://www.zerogeography.net/2014/04/my-response-to-geoweb-and-big-data.html (http://www.zerogeography.net/2014/04/my-response-to-geoweb-and-big-data.html)Accessed May 10, 2014.

Hahmann S and D Burghardt 2013 How Much Information is Geospatially Referenced? Networks and Cognition. *International Journal of Geographical Information Science 27*(6): 1171-1189.

Haklay M 2013 Neogeography and the Delusion of Democratization. *Environment and Planning A 45*(1): 55-69.

Haklay M and P Weber 2008 OpenStreetMap: User-generated Street Maps. *Pervasive Computing, IEEE 7*(4): 12-18.

Haraway D 1991 *Simians, Cyborgs, and Women: The Reinvention of Nature*. New York: Routledge.

Harford T 2014 Big Data: Are We Making a Big Mistake? *Financial Times Magazine.* http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2ysdIXgD2 (http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2ysdIXgD2) Accessed May 1, 2014.

Harman G 2010 Technology, Objects and Things in Heidegger. *Cambridge Journal of Economics 34:* 17-25.

Harris L and H Hazen 2005 Power of Maps: (Counter) Mapping for Conservation. *ACME 4*(1): 99-130.

Hidalgo CA 2014 Saving Big Data from Big Mouths. *Scientific American*. http://www.scientificamerican.com/article/saving-big-data-from-big-mouths/ (http://www.scientificamerican.com/article/saving-big-data-from-big-mouths/) Accessed May 1, 2014.

Holmes B 2003 Imaginary Maps, Global Solidarities. *Piet Zwart Institute*.

Humphreys L, G Phillipa, and B Krishnamurthy 2014 Twitter: A Content Analysis of Personal Information. *Information, Communication & Society 17*(7) 843-857.

K.N.C. 2014 The Backlash Against Big Data. *The Economist*. http://www.economist.com/blogs/economist-explains/2014/04/economist-explains-10 (http://www.economist.com/blogs/economist-explains/2014/04/economist-explains-10) Accessed May 10, 2014.

Kitchin R 2014 Short Presentation on the Need for Critical Data Studies. *The Programmable City* blog. http://www.nuim.ie/progcity/2014/04/short-presentation-on-the-need-for-critical-data-studies/ (http://www.nuim.ie/progcity/2014/04/short-presentation-on-the-need-for-critical-data-studies/) Accessed May 10, 2014.

Kitchin R and T Lauriault 2014 Small Data, Data Infrastructure and Big Data. The Programmable City Working Paper 1. http://dx.doi.org/10.2139/ssrn.2376148 (http://dx.doi.org/10.2139/ssrn.2376148) Accessed May 10, 2014.

Kitchin R and M Dodge 2011 *Code/Space: Software and Everyday Life.* Cambridge MA: The MIT Press.

Latour B 1988 *How to Follow Scientists and Engineers Through Society*. Harvard University Press.

Leyshon A, S French, N Thrift, L Crew, and P Webb 2005 Accounting for E-commerce: Abstractions, Virtualism, and the Cultural Circuit of Capital. *Economy and Society 34*(3): 428-450.

Marcus G and E Davis 2014 Eight (No, Nine!) Problems with Big Data. *The New York Times*. http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?_r=3 (http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?_r=3) Accessed May 1, 2014.

Marcus G and E Davis 2013 What Nate Silver Gets Wrong. *The New Yorker*. http://www.newyorker.com/online/blogs/books/2013/01/what-nate-silver-gets-wrong.html (http://www.newyorker.com/online/blogs/books/2013/01/what-nate-silver-gets-wrong.html) Accessed May 9, 2014.

Massey D 1993 Power-Geometry and a Progressive Sense of Place. In: J Bird, B Curtis, T Putnam, G Robertson, and L Tickner (eds) *Mapping the futures: Local cultures, global change*. New York: Routledge 59-69.

Mayer-Schoernberger V and K Cukier 2013 *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* New York: Houghton Mifflin Harcourt.

Merrifield A 1995 Situated Knowledge through Exploration: Reflections on Bunge's "Geographical Expeditions." *Antipode 27*(1): 49-70.

Mogel L and A Bhagat (eds) 2007 *An Atlas of Radical Cartography*. Los Angeles: Journal of Aesthetics & Protest Press.

Pavlovskaya M 2006 Theorizing with GIS: A Tool for Critical Geographies? *Environment and Planning A 38*. 2003-2020.

Pearce R 2013 Big Data is BS: Obama Campaign CTO. *CIO*. http://www.cio.com.au/article/462961/big_data_bs_obama_campaign_cto/ (http://www.cio.com.au/article/462961/big_data_bs_obama_campaign_cto/)Accessed May 1, 2014.

Peluso NL 1995 Whose Woods Are These? Counter-Mapping Forest Territories in Kalimantan, Indonesia. *Antipode 27*(4).

Pickles J 1997 Tool or Science? GIS, Technoscience, and the Theoretical Turn. *Annals of the Association of American Geographers 87*(2): 363-372.

Podesta J 2014 Remarks Delivered by Counselor John Podesta, The White House OSTP. Presentation at the Berkeley Center for Law and Technology Workshop: "Big Data: Values and governance." UC Berkeley School of Information, Berkeley, CA, April 1, 2014. http://www.whitehouse.gov/sites/default/files/docs/040114_remarks_john_podesta_big_data_1.pdf (http://www.whitehouse.gov/sites/default/files/docs/040114_remarks_john_podesta_big_data_1.pdf) Accessed May 11, 2014.

Schulten S 2012 *Mapping the Nation: History and Cartography in Nineteenth-Century America.* Chicago IL: University of Chicago Press.

Shah V 2013 Map of Stop and Frisks in NYC in New York City Show Concentration by Race and Neighborhood. http://untappedcities.com/2013/08/28/new-map-shows-police-stop-and-frisks-according-to-race-and-neighbourhood-in-new-york-city/ (http://untappedcities.com/2013/08/28/new-map-shows-police-stop-and-frisks-according-to-race-and-neighbourhood-in-new-york-city/)Accessed May 14, 2014.

Sieber R 2006 Public Participation Geographic Information Systems: A Literature Review and Framework. *Annals of the Association of American Geographers 96*(3): 491-507.

Terranova T 2004 *Network Culture: Politics for the Information Age*. London: Pluto Press.

Thatcher J Forthcoming (2014) Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data. *International Journal of Communication*.

Thatcher J and R Burns (organizers) 2013 Whither Small Data?: The Limits of "Big Data" and the Value of "Small Data" Studies. Session at the National Meeting of the Association of American Geographers, Los Angeles, CA, April 12, 2013. http://meridian.aag.org/callforpapers/program/SessionDetail.cfm?SessionID=17195 (http://meridian.aag.org/callforpapers/program/SessionDetail.cfm?SessionID=17195) Accessed May 15, 2014.

Tsou MH 2014 Building a New Research Agenda for Geographers: Human Dynamics in the Mobile Age (HDMA). Presentation at the National Meeting of the Association of American Geographers, Tampa, FL, April 10th, 2014. Quoted from the abstract. http://meridian.aag.org/callforpapers/program/AbstractDetail.cfm?AbstractID=55574 (http://meridian.aag.org/callforpapers/program/AbstractDetail.cfm?AbstractID=55574) Accessed May 2, 2014.

Tsou MH, IH Kim, S Wandersee, D Lusher, L An, B Spitzberg, D Gupta, JM Gawron, J Smith, JA Yang, and SY Han 2014 Mapping Ideas from Cyberspace to Realspace: Visualizing the Spatial Context of Keywords from Web Page Search Results. *International Journal of Digital Earth 7*(4): 316-335.

Tuan YF 1991 Language and the Making of Place: A Narrative-Descriptive Approach. *Annals of the Association of American Geographers 81*(4): 684-696.

Vaidhyanathan S 2006 Afterword: Critical Information Studies: A Bibliographic Manifesto. *Cultural Studies 20* (2-3): 292-315.

Wainwright J and J Bryan 2009 Cartography, Territory, Property: Postcolonial Reflections on Indigenous Counter-Mapping in Nicaragua and Belize. *Cultural Geographies 16*(2): 153-178.

Wilson MW 2014 Morgan Freeman is Dead and Other Big Data Stories. *Culture Geographies*. DOI: 10.1177/1474474014525055.

Wood D 2010 *Rethinking the Power of Maps*. New York: Guilford Press.

Wright D, M Goodchild, and J Proctor 1997 Still Hoping to Turn That Theoretical Corner. *Annals of the Association of American Geographers 87*(2): 373.

Yeager C and T Steiger 2013 Applied Geography in a Digital Age: The Case for Mixed Methods. *Applied Geography 39*(May): 1-4.

Zook M and A Poorthius 2014 Offline Brews and Online Views: Exploring the Geography of Beer Tweets. In: M Patterson and N Hoalst-Pullen (eds) *The Geography of Beer*. http://link.springer.com/chapter/10.1007/978-94-007-7787-3_17# (http://link.springer.com/chapter/10.1007/978-94-007-7787-3_17)Accessed May 14, 2014.

16 thoughts on "Craig Dalton and Jim Thatcher – What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'"

1. *Surveillance & Society new issue on Big Data Surveillance | Society and Space - Environment and Planning D* says:
   May 19, 2014 at 12:53 pm

Article

new media & society
1–17
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1461444816629470
nms.sagepub.com
SAGE

Hacking in the public interest: Authority, legitimacy, means, and ends

Alison Powell
London School of Economics and Political Science, UK

Abstract
The cultural appropriation of ideas about hacking and opening knowledge have had significant impact on ways of developing participation in creating public interest knowledge and knowledge commons. In particular, the ideal of hacking as developed through studies of free and open source (F/OS) has highlighted the value of processes of participation, including participatory governance, in relation to the value of expanded accessibility of knowledge, including knowledge commons. Yet, these means and ends are often conflated. This article employs three examples of projects where hacker-inspired perspectives on scientific knowledge conflict with institutional perspectives. Each example develops differently the relationships between means and ends in relation to authority and legitimacy. The article's analysis suggests that while hacker culture's focus on authority through participation has had great traction in business and in public interest science, this may come limit the contribution to knowledge in the public interest - especially knowledge commons.

Keywords
Authority, hacking, knowledge construction, open science, open source, public interest

Introduction

In trying to understand the cultural significance of hacking and "do it yourself" (DIY) culture, it is easy to conflate means and ends. Much research on hacking has focused on

Corresponding author:
Alison Powell, Department of Media and Communications, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK.
Email: a.powell@lse.ac.uk

how participatory engagement creates alternative ways of engaging with machines, intellectual property, or material. At the same time, there is another interpretation threaded through scholarship on hacking—that it has ends, and that the hack might transform the way the machine works, the way knowledge is shared, or the material object's final form. This interpretation is especially salient for studies of hacking in the free software tradition and in discussions of the importance of creating knowledge commons using intellectual property hacks. Yet, means and ends are rarely separated, even when hacking culture is explicitly connected with the notion of the public interest, as it is in relation to open knowledge and science. This article pursues two (slightly contradictory) goals: (1) to advance an ethical critique of the focus on participation within hacking culture and (2) to explore how this focus on participation conflates the means and ends of hacking practice, possibly to the detriment of knowledge commons and to radical democratization of scientific practice. To draw out the first point, it builds upon Mansell's (2013) analysis of how modes of authority become significant for managing knowledge commons, examining the relationship between the legitimacy of participation within hacking culture and "adaptive" forms of authority. To develop the second, it extends Collins and Evan's (2002) discussion of expertise and public knowledge and reveals how a focus on authority and legitimacy in relation to participation (rather than engagement with other politics of expertise) prevent a true focus on what the outcomes of hacking might produce for expanded notions of the public interest and knowledge commons.

To develop these two linked arguments, I consider three examples of projects where hacking culture is positioned as contributing to the public interest and knowledge commons. In the first example of the development of the CERN Open Hardware License (CERN OHL), researchers in the Beams and the Knowledge Exchange Sections at the European High-Energy established an open source community that deliberately included members whose authority emerged from their long association with either open source license development or the practice of open hardware development. In the second example, the non-profit Public Lab employs open hardware as part of a strategy for broadening environmental inquiry that is consciously linked to DIY ethics and what Ratto and Boler (2014) refer to as "critical making." In this mode, the DIY ethos is a "critical" activity that "provides both the possibility to intervene substantively in systems of authority and power and that offers an important site for reflecting on how such power is constituted by infrastructures, institutions, communities and practices" (p. 1). In the third example of the Internet of Things Academy (IoTA), more accessible hardware raises questions about what kinds of scientific data garner more legitimacy. Designers on this project employ hardware sensors including noise and air quality monitors that produce well-calibrated measurements of similar quality to those used by scientific professionals. All of the examples engage with the notion of open hardware, enrolling these projects in debates about the means and ends of hacker projects like the General Public License (GPL; see Powell, 2012).

Open hardware raises questions about how to extend the provisions for keeping intellectual property in commons. This is similar to what Barron (2013) refers to as "the tendency to problematize the technical infrastructures underpinning today's digitally mediated public spheres" (p. 599). To practitioners seeking to maximize participation in technological or scientific knowledge production, open hardware puts equipment in the hands of ordinary people, permitting a hack of science itself (McQuillan, 2014). In addition, open hardware can contribute to a DIY ethic of creative repurposing that positions hacking, tinkering, and making of scientific measurement equipment as intrinsically political. The examples in the article illustrate how conflicts of authority within hacking culture gain greater political significance when they are played out over concerns related to the public interest and knowledge commons.

Hacking culture and its conflicts

Hacking and hackers have transformed the social world outside of software. Social researchers situate hacking as a form of spontaneous techno-cultural jouissance (Jordan, 2008; Levy, 1984; Thomas, 2002), as a model for participation-based governance (Dafermos, 2012; Kostakis et al, 2015; Mateos-Garcia and Steinmueller, 2008) with the possibility to transform markets more broadly (Benkler, 2006, 2011), as the enactment of critiques of the politics of technological systems (Barron, 2013; Bodó, 2014; Kelty, 2005; Sauter, 2013) and intellectual property systems (Barron, 2012; Lessig, 2006), or as an engagement with the culture and performance of masculinity and expertise (Dunbar-Hester, 2010). We also assess the relationship between hacking and the social, political, and economic systems that are transformed by expansions of hacking practice. When Wark (2013) writes that "the hacker makes something new out of property that belongs to everyone in the first place" (p. 73), he, like Soderberg (2008), claims that hacking reveals as well as transforms cultural and technical products, making us aware of their status as common knowledge usable by all.

Here, we are reminded that hacker participation in creating projects like free and open source software (F/OS) led to the development and transformation of political values like liberalism (Coleman, 2012) through the development of "recursive publics" who create the means for their own perpetuation (Kelty, 2008) and through the reworking of intellectual property regimes to suggest possibilities for the creation of knowledge in commons (Lessig, 2006; Stallman, 1999). We are also reminded of the possibilities for "democratic rationalization" of technology (Feenberg, 2008; Milberry, 2014) and hence the possibility that hacking, as a form of DIY practice, could prefigure or perhaps exemplify a democratization of technical or scientific knowledge. These varying political outcomes also include the contention that participation in hacking and other DIY projects models a democratization of knowledge through "critical making" (Ratto, 2011; Ratto and Boler, 2014) and a potential transformation and broadening of scientific publics through DIY speculation through design (DiSalvo, 2014).

At the same time, features of hacking culture can re-invigorate existing cultural formations, for example, through the development of open source organizational culture within F/OS and its subsequent embedding of participation-based value within software production economies (Berdou, 2011; Weber, 2004), and "prosumer" practices (Moody, 2002) leading to the development of different products (Von Hippel, 2005). Also in the economic sphere, the participation and networked relationships have been claimed as foundations for a network and reputation-based economy (Benkler, 2006). This reading of hacking culture celebrates individualism, participation, and reputation within a "new spirit" of capitalism (Barron, 2013) and neoliberal governance (Cammaerts, 2011).

In this review of hacking's significance, two strands emerge: one, a valorization of participation, both as a feature of governance and as a mode of engagement with institutional power and, two, an evocation of a transformation of knowledge production and accessibility, extending from technical to scientific knowledge. The two strands illustrate how hacking culture is associated with transformations in means (participation) and ends (most often, the modes of production of software, but now, the modes of production of scientific and public interest knowledge). These are often conflated. This article intervenes in this debate to provide an analysis of the consequences of the focus on participation rather than outcomes of hacking. These consequences include the market appropriations of hacking processes already considered in the literature (Cammaerts, 2011; Powell, 2012) as well as a limited transformation of the processes of scientific and public interest knowledge production.

The article builds on previous work on authority and legitimacy in relation to both participation and knowledge production in hacking culture, especially the principles of adaptive and constituted authority developed by Mansell (2013), and the assessment of how contests of legitimacy (Collins, 2010) relate to transformations in knowledge practices. This extends previous work on governance and participation within hacker communities (Kostakis, 2010; Kostakis et al., 2015), particularly F/OS hackers (Dafermos, 2001; Dafermos and Söderberg, 2009), but also follows a turn in the science studies' literature on expertise that has become increasingly concerned with how expertise is legitimated in different contexts.

Authority and legitimacy: F/OS and the GPL

Hacking culture sets up novel dynamics of authority: hackers are understood to establish their own authority, or "knowledge of purpose and technique acquired and demonstrated through participation" (Mateos-Garcia and Steinmueller, 2008: 336). In contrast to the authority associated with institutions accrued through symbolic reinforcement of the functional necessity for an institution (Castoriadis, 1987 [1975]), the authority associated with hacker culture is rooted in the imagination of participation and in expertise consolidated through participation. Other scholars of hacking in the DIY vein have focused on how participation in building and rebuilding technology operate as strategies for eroding boundaries between experts and laypeople and redistributing authority (Dunbar-Hester, 2014).

These forms of authority and legitimacy have also supported existing institutions, especially the institution of market capitalism. Much work over the past decade has identified how hacking practices; especially, those related to F/OS production contribute to expertise and economic production within firms (Mansell and Berdou, 2010; Tapscott and Williams, 2006; Weber, 2004). Barron (2013) notes that the valorization of individual participation that is part of F/OS production has significant risks for the notion of collective goods:

> In a reputational economy, creative production becomes a means to the end of forging a publicly recognizable identity: the goal is not so much to produce a body of work that can take on an existence beyond oneself as to become an entrepreneur of one's self by associating as much

> activity as possible (preferably including that of others) with one's name. If unchecked, this will yield a culture in which (self-) promotion takes priority over production; it is also liable to obscure the collective effort that sustains every project, erode mutual trust and loyalty, and ultimately undermine the FOSS spirit itself. (p. 618)

Barron identifies that the relationship between the means of authority developed through participation and the ends of production and collective value are collapsed and obscured by some features of the development of authority through participation. Other work goes further to examine the ambiguity of authority in relation to both participation and possible ways to develop (or value) knowledge. This second set of ideas raises questions about whether the emphasis on participation in hacking culture has consequences for its role in democratizing scientific knowledge production.

Contests of authority

In Mansell's (2013) analysis of how modes of authority become significant for managing knowledge commons, and in Collins and Evan's (2002) discussion of expertise and public knowledge, researchers identify how legitimacy develops through processes of participation which may not be matched in which ways they are perceived as being resolved.

Adaptive knowledge legitimated through participation. Mansell (2013) outlines how modes of authority become significant for managing knowledge commons, exploring the potential for collaborations between formal science professionals and loosely organized groups of people working on crowdsourcing projects. Differences in data curation highlight differences in the nature of authority—along a continuum between the "constituted authority" of hierarchical relations established in reference to formal norms of science and its institutions, and the "adaptive authority" "characterizing loose, bottom-up, often informal, forms of authority that are frequently associated with information activities of many loosely connected online groups" (Mansell, 2013: 256). Within these specific communities of practice, different individuals collect data that are valued differently depending on the form of authority the person's associated with. Constituted authority validates the participation of the individuals who are part of the crowd. Adaptive authority validates the quality of the data and its later utility for scientific practice. While the practices associated with adaptive authority valorize the aggregation and sifting of knowledge for immediate use (such as collecting information in online repositories), constituted authority is concerned with curation of "useful" scientific information and validation of who gets access to that knowledge.

The notions of constituted and adaptive authority are helpful in developing a response to the challenge of creating "open innovation" in scientific practice. In particular, Mansell's distinction between forms of authority highlights how scientific expertise remains associated with the valorization of certain forms of knowledge and control of their access. In scientific crowdsourcing, people outside of scientific institutions more often value knowledge for its immediate purpose, or for the reputational value that contributing might garner. This conflicts with forms of constituted authority that are more closely associated with "expert" knowledge located within scientific institutions. The crowdsourcing dynamics that are the subject of Mansell's inquiry often create a power imbalance whereby "lay" contributors to crowdsourced scientific projects are positioned as amateurs and as data sources, rather than as collaborators.

This kind of contest between adaptive and constituted authority mirrors the kinds of contests usually associated with hacker culture, in which hacker ethics of critique and revelation are placed in contrast with ethics of enclosure.

Here, in addition to the contested politics of authority, two further dynamics emerge: a politics of expertise, which distinguishes expert and lay knowledge and which aligns with participation as the means of hacking culture, and a politics of commoning, which seeks to connect them through the development and management of knowledge commons—arguably the desirable ends for public interest hacking culture.

The politics of expertise. Collins and Evans (2002) highlight how expertise must be identified for political ends. They note that even within a framework where multiple forms of expertise are valuable, some gain greater legitimacy. There is one kind of expertise, often scientific, that "has gained a kind of universal authority across society in virtue of what everyone believes to be its efficacy" (p. 251). In relation to this expertise others emerge, including a type of "contributory expertise" that is in a field relevant to this highly legitimate expertise. Judgments then need to be made about the legitimacy of contributory expertise. Collins and Evans write,

> it is more difficult to separate the credentialed scientist from the experienced practitioner than was once thought: when we move toward experience as a criterion of expertise the boundary around science softens, and the set of activities known as "science" merges into expertise in general. (p. 253)

In this context, what becomes important is not expertise but legitimacy. Legitimacy can be conferred through relationships to structures of authority but also—as all of the previous studies of hacking culture identify—through resistance to structures of authority. This hinge point between authority and legitimacy motivates interest in expanding access to a creation of scientific knowledge: as Collins and Evans point out, the high levels of legitimacy associated with "core" scientific legitimacy lead to lower levels of certainty. This in turn means that other actors play roles in conferring legitimacy: the media, the people with "contributory expertise," and other people in general. In this context, efforts to "democratize science" in the public interest can be viewed as sites for the negotiation of legitimacy—the kind of sites we will shortly be discussing in relation to hacker culture and public interest science.

The democratization of scientific production is considered through research on "open science." The concerns of open science often have to do with the capacity to provide broader access to the literature, experimental materials, and data sharing (Wilbanks, 2007), or the capacity to integrate different types of information and knowledge as part of a broader innovation process (Lakhani and Panetta, 2007). These concerns foreground "openness" related to accessibility, whether of code, data, or knowledge. This contrasts with the research on free and open source software (F/OSS; Coleman, 2012) that places an emphasis on the process and politics of opening things up, where "openness" connects with a politics of critique. They are also implicitly oriented toward participation as a value in itself rather than in orientation to an outcome, but this literature, more than the more canonical discussions of hacker culture's governance processes, also gestures toward the ideal outcome of the knowledge commons

The politics of knowledge commons. The institutional arrangement of maintaining resources in commons has been thoroughly investigated by Ostrom (1990) and expanded through studies on various forms of commons data management (Fuster Morell, 2010) and open source software. Much previous work on the institutional management of such "knowledge commons" has investigated forms of social ordering and governance (Madison et al., 2014) generating important insights on how commons can be maintained. The commons has an orthogonal relationship to hacker culture. It is not necessarily always the end goal of hacking, in contrast to the expression of individual liberty that Coleman and Golub (2008) link to practices of F/OS activism. In the economic realm, a similar legitimacy linking participation to the "spirit of capitalism" (Boltanski and Chiapello, 2005 [1999]) has become integrated into "lean" "networked" modes of production. This, as Barron (2013) points out, "embed[s] new modalities of control over both production and consumption, and extend[s] commodification processes rather than curtailing them" (p. 609). The question then becomes, as Mansell's work reiterates, whether the kinds of authority associated with "contributory expertise" and networked participation of the kind valorized within hacking culture are able to transform other structures of power rather than being subsumed to them. F/OS production does create a knowledge commons of re-usable intellectual property, and the extension of this commons was one rationale for the development of open source hardware licenses (see Powell, 2012).

As an organizing and political principle, the commons challenges some of the separations between forms of expertise as outlined by Collins and Evans and evokes the promise of hacking to enact disruption to release to the people something that always should have been liberated (to paraphrase Wark, 2004). The following section tracks how this promise has been enacted through different types of participation across three projects linked by their engagement with open hardware in relation to science and the public interest. They illustrate that contests of authority in relation to legitimacy often play out as confusions between the means and ends of "hacking" knowledge systems.

Examples

CERN OHL

The first example, of the development of the CERN OHL, directly draws upon the notion of the knowledge commons as a means of integrating knowledge drawn from hacker and advocacy communities with knowledge produced within CERN. It also shows how hacker culture animates this integration, providing a way to highlight the flexibility and openness of a particular group of CERN researchers. The development of the CERN OHL thus fits within a longer history of knowledge exchange at the institute, but seeks a different kind of engagement with the "non-expert" partners than some other projects.

Since its founding in 1954, the European high-energy physics lab (CERN) has intentionally developed strategies for intensive scientific collaboration (Krige, 2001). The center is associated with what Collins (1998) has identified as "open evidential culture." CERN's most recent, complex, and multidisciplinary work, the creation of the ATLAS particle detector and the development of the Large Hadron Collider, have also required intense collaboration employing distributed working processes that brought together culturally heterogeneous researchers working in very different institutional settings (Boisot, 2011). Boisot's description of the work on the ATLAS detector draws on a narrative familiar to scholars of F/OS and open collaboration, highlighting collaboration and "emergent strategies" that Boisot frames as typical of adhocracy (Toffler, 1970). In his report, the flexibility of the membership's work was coordinated around the detector, solidified by shared values among the many participants, and facilitated by the use of information and communication technologies (ICTs). This narrative of flexibility and collaboration has been part of the institutional identity of CERN (see Collins, 1998; Knorr Cetina, 1999), supporting the efforts to develop open hardware as a means to foster collaboration within and outside the institution. Although CERN uses crowdsourced science projects as one of their knowledge transfer strategies, the CERN OHL project is unique in that its public engagement comes mainly through the process of defining the parameters of the open source project.

Javier Serrano of the Beams Section and Myriam Ayass of the Knowledge Transfer section launched the project in 2011 as a way to standardize the intellectual property relations of submission to the repository for open hardware designs that Serrano had developed. In 2011, the two published a first version of the license and began consultation with hardware hackers and other open hardware advocates, visiting open hardware conferences and Maker Faires and establishing a public mailing list. The expertise and experience of the participants in these conferences and mailing list discussions was perceived as being essential for the development of the license.

The license was intended to provide a parallel for electronics designs to the GPL licensing that applies to all software code written at CERN. The GPL was chosen because

> Open Source principles encourage the creation of open communities and collaborations of users invited to improve and complement the software and share their enhancements with the entire community. This accords with the historical CERN collaborative spirit and maximizes the in-kind return to CERN. In substance, this recommendation promotes the concept of collaborative dissemination … the Copyleft philosophy fits best with CERN scientific philosophy and tradition. (Fluckiger, 2012)

The support within CERN for "collaborative dissemination" foreshadows some of the challenges between balancing the means through which software is improved by greater numbers of participants and the ends to which "collaborative dissemination" might be directed.

In interviews with Serrano and with members of the Knowledge Exchange team, it is clear that the license was developed in order to facilitate collaboration with companies and with individual hackers and makers. But the interest was not only in ends, that is, in having a final product that would allow the Beams Section to work more closely with companies, or benefit from discussions among open source advocates. It was also about means, and the significance of employing a process that respected the expertise outside of CERN as much as inside. In this process, hardware hacker and advocacy communities were positioned as essential to the development of the open hardware license: “I see it sometimes as enlarging our team … because the documents are all public, if [a collaborator] happens to be from a company … he’s just one more guy collaborating” (Javier Serrano, 2013, personal communication). Javier Serrano (2013, personal communication) describes himself as a facilitator but insists that he is not skilled enough to be a hacker:

> I know impressive hackers, and I would not say I am in the same league as them. But I believe in teams a lot, and what I am doing allows them to do very cool stuff, so I’m very proud of that.

This vision is of an integrated team, where both the “impressive hackers” located outside of CERN and the researchers within can work toward the same goal. In the CERN OHL project, the goal was to create a hardware license in the same mode as the GPL. This was for two reasons: Serrano was himself a free software advocate and inspired by the notion of creating an ever expanding knowledge commons that would include hardware as well as software. His efforts to establish the CERN OHL contributed to an ecosystem of open hardware licenses that reflected different philosophical and political stances (see Powell, 2012). Gaining legitimacy and support from the open source advocacy community and from hardware hackers was essential for Serrano’s broader goal of extending the GPL into new contexts.

To gain this legitimacy, the license was discussed on a mailing list. As Powell (2015) describes, these discussions demonstrated the difficulty of determining what open hardware referred to: accessible designs, plans whose components are totally re-usable, a better form of knowledge commons, or recipes and descriptions for construction placed in a repository. Contention between these different ways of thinking about open hardware was in part resolved by allowing the license to act as a “boundary object” (Star, 2010)—a shared framework that permits collaboration between groups developing different kinds of knowledge.

The resolution of the mailing list discussion solidified the importance of adaptive knowledge and hacker relationships for the CERN OHL. As a result of the discussion, the license’s new version included provisions that favored the open source community's interests over those of the Beams Section and the Knowledge Exchange Section. The new version of the license removed a provision that would require anyone who used a licensed design to inform the person who licensed it about how they were using it. This would have been very helpful for CERN, since it would have allowed the Knowledge Exchange section to monitor the use of material and ideas developed within CERN. Instead, removing this clause valorized the interests of the open source community participants and aligned with their adaptive authority. It developed the means of collaboration rather than the ends of better identifying open source materials.

Not all members of the open source community supported the development of a GPL-type license for hardware as the best way to create and broaden a knowledge commons related to electronic designs. Longtime open source advocate Bruce Perens, one of the participants in the CERN OHL license development, and a well-known developer of open software licenses and standards argued that open hardware licenses have the unintended consequence of creating more, not less, focus on intellectual property. This grates against the hacker perspective on these issues. In an interview hosted on hacker site Slashdot, Perens writes,

> There's an important thing we should be aware of about Open Hardware. It's backwards in a way. Richard Stallman's Free Software movement opposed software being copyrighted. Copyright does not, for the most part, apply to hardware designs because they are functional … Patents apply to hardware designs, but most Open Hardware designers never pursue a patent on their designs. What then do they license to others?
>
> It turns out that we have a group of people at CERN, and one of my favorite lawyers and Yahoo, and even me, trying to add restrictions to something that is, for the most part, already in the public domain. And it came to me that this was backwards, and that we could be working against our own interest that way … The problem is that when we start licensing things that are actually in the public domain, we create norms that the courts take seriously … If we were responsible for taking hardware designs from public domain to copyrighted status, we'd be shooting ourselves in the foot. (Perens, 2014)

Perens is concerned that the efforts at resisting enclosure of intellectual property and continuing to allow space for critique of these frameworks is actually being limited by the move toward licensing. He is concerned that focusing on means and valorizing adaptive authority might limit the positive consequences of hacking by rendering much knowledge inaccessible—a fundamental impediment to facilitating further re-use of common resources, and perhaps a brake on hacking practice.

The development of the CERN OHL, then, is a consolidation of a particular perspective on the extension of GPL-inspired legal frameworks. In the extensive discussions on the CERN OHL mailing list analyzed in Powell (2015), the challenge of successfully extending the principles of the GPL so that they would fully apply to electronics did not quite overlap with the interest in employing GPL principles to either expand a knowledge commons or to monitor CERN's intellectual property. As Perens' critique highlights, participation in modifying the license, and valorization of that participation against authority of CERN, inadvertently valorizes a narrower interpretation of open hardware and may even have the consequence of limiting the expansion of open hardware knowledge commons. This illustrates the long-term consequences of valorizing participation for its own sake and highlights the tensions between adaptive and constituted authority.

Public Lab

In the second example, the US non-profit Public Lab also engages with ideas of open hardware and hacker cultural ethics, this time in relation to the democratic ethics of DIY. Public Lab, a non-profit organization based in the United States but with local projects running in locations around the world, develops and applies open source tools to environmental exploration and investigation. With an explicit focus on democratization of scientific knowledge through making, the project came to prominence after it used home-made balloons and digital cameras to map the Gulf oil spill in 2010. It aims at breaking down inequities of knowledge production by supporting DIY methods of collecting scientific data:

> DIY aims to make technology something anyone can develop; PublicLab aims to make scientific research in environmental issues something anyone can do well. To make something oneself is to have a sense of ownership of it, and we extend this sense to scientific tools and data. (Warren and Regalado, 2014: n.p.)

Public Lab runs workshops around the world that teach people how to build relatively low-cost tools for environmental monitoring and community mapping, including kite-mounted digital cameras. Cindy Regalado (2014, personal communication), a London-based member of Public Lab, explains that these projects are intended to develop a "spark of interest" among people and to employ DIY methods to help them understand that they could make their own monitoring tools to use in any kind of project. For Public Lab open source is understood as an ethic, linked to the DIY ethic of creative repurposing of objects. The project aims to democratize scientific inquiry by democratizing the production of its measurement tools, but more specifically to expand the ability of people to feel capable of pursuing an interest or curiosity.

Public Lab's interpretation of open source aligns with a different politics of expertise than the integration of "contributory expert" authority to knowledge sharing at CERN. For Public Lab, the ethic of open source that motivates their projects is concerned with increasing accessibility of knowledge and allowing more people to understand how to collect and represent information about their lives and communities. In this enactment of public interest science, the public interest is served by the public understanding the principles of science and feeling empowered to participate. Although the project is best known for supporting local residents in designing and deploying home-made aerial cameras to map the local impact of the Gulf oil spill, advocates stress that the purpose of these projects is not to develop tools that produce scientifically verifiable data, but rather to encourage participation in creating tools and understanding science.

This is especially evident in PLOTS (or The Public Laboratory for Open Technology and Science), Public Lab's open knowledge repository, which includes research notes, designs, and instructions on how to build scientific measurement tools, including aerial cameras assembled from inexpensive digital cameras and large home-made kites. While some electronics designs shared on PLOTS use the CERN OHL, the repository is mostly meant to allow people to openly share, create, and reproduce tools for measurement and story telling. The knowledge is "open" because the equipment is relatively inexpensive and because know-how is shared through the research notes and instructions.

PLOTS valorizes adaptive knowledge. It focuses on the financial accessibility of materials and the significance of participation in using them and does not necessarily collect or share the results of that participation. It decenters scientific value away from sites of constituted knowledge and authority, which place more value on the quality of scientific results. Public Lab grounds knowledge in material practice—as their 2013 annual report reads, "creating tools and communities of expertise (whether local or scientific)" (PublicLab, 2014). While this has significant value as a way of valorizing alternatives to constituted authority, it also reinforces a divide between modes of authority, where scientific institutions are still sources of important knowledge, but not necessarily collaborators in the horizontal processes of co-creation. Furthermore, there is an important difference in how open hardware is imagined in the CERN-OHL and in the Public Labs contexts. In the former, open hardware refers to design specifications sufficient to allow the electronics to be constructed by someone with the appropriate skills, in the latter, to financial accessibility and ease of construction. These two different ways of conceiving of open hardware do align, as open source designs that can be re-used make hardware like the Arduino lower in cost and easier to use. But they also diverge. Attempts like the CERN OHL to develop a stock of re-usable hardware designs through the integration of hacker practices into scientific collaboration imagine open hardware differently than the Public Labs projects that valorize knowing through making.

As with the CERN case, there are complexities that highlight the differences in legitimacy in relation to means, and legitimacy in relation to ends. The DIY objects constructed in Public Lab projects help people without scientific knowledge to develop and amplify their comfort with scientific practice. However, this positions scientific knowledge and authority as something separate, rather than as something to be collectively developed. In terms of process, this means that the opportunities for consistent negotiation between forms of authority are more limited. In terms of result, the separation between forms of authority widens, and the legitimacy of institutional science is reinforced by the fact that the data collected by inexpensive sensors are often of poor quality or not comparable with data produced by scientific institutions. This distinction is at the heart of the separation that Mansell identifies between the two forms of authority. As she notes, this separation complicates efforts at establishing knowledge commons because of the conflict between different perspectives on which kinds of knowledge ought to be part of such commons. Finally, the explicit association between material engagement and empowerment, while central to the mobilization of hacking culture, also reveals the fractured relationships between technical prowess and other forms of empowerment related to race and gender (Dunbar-Hester, 2010). For Public Lab, shareable knowledge is not an end goal, but part of the process that is intimately linked to making and doing. All of the legitimacy is thus associated with means, rather than with ends that could include an ongoing scientific conversation or the production of scientific data.

IoTA

The final example, the IoTA run by the Superflux (2014) design agency, more accessible hardware raises questions about what kinds of scientific data garner more legitimacy. Designers on this project employ environmental sensors including noise and air-quality monitors that produce well-calibrated measurements of similar quality to those used by scientific professionals including policy-makers. Data from these sensors are intended to challenge government data with data collected by citizens with particular concerns (aircraft noise and air quality). The quality of data (and thus of the hardware) becomes more important than their accessibility to the citizens.

IoTA has two pilot projects that use sensor-based networks (the "Internet of Things") to address civic concerns. These are designed so that engagement with the design of data collection and analysis is very accessible, while not insisting that participants must engage in construction of hardware. The IoTA project is meant to help to valorize things that citizens already know about, by employing sensor technologies along with "little data" collection technologies like daily notebooks. The first pilot called NoiseNap, measured noise pollution under London flightpaths, and BuggyAir, a project currently under development, will distribute air-quality sensors to families to mount on their baby buggies. These sensors will then measure air quality as it is experienced at ground level and in areas where children are traveling.

The BuggyAir project in particular encourages the development of very high quality data, according to Superflux founder Anab Jain. This is to encourage two possible outcomes: first, behavior change in participants and other individuals as a result of the BuggyAir readings (this might include avoiding walking on routes where the sensors record very high air pollution) and, second, policy change on the part of governments and standards setters who might respond to legitimate high-quality data. Jain (2015, personal communication) explains,

> Quality is important. How can you have accurate enough data so you can advocate for car companies to consider new standards for brakes [that are one of the major contributors to particulate matter (PM) ground level air pollution]. This is small data. It will never be big data, so it has to be good data.

In contrast to the approach of Public Lab, where financial accessibility of hardware is a key feature of the project's openness and accessibility, BuggyAir employs proprietary sensors that cost £500 each and which are precisely calibrated to have 97% accuracy in measuring air pollution of all types, including particulate matter which composes 80% of ground level air pollution in London. This calibration and quality are understood as increasing the legitimacy of citizen-collected data. Jain (2015, personal communication) and her team are concerned that the very accessibility of inexpensive scientific tools may mean that the data they produce is not considered legitimate from the perspective of constituted authority: "these citizen science projects, they might have a button you can wear, but the data is not even 50% reliable."

The IoTA pilots stress the legitimacy of their sensor data as a pathway toward valorizing citizen perspectives. In the NoiseNap pilot, the sensor data on noise levels are placed together with journal entries describing the context and experience of aircraft noise. However, in both pilots, the technologies of scientific measurement are black boxed. Thus, the projects valorize non-expert knowledge and the adaptive authority that investigate its social and economic context, but do so by closing off the data collection and making its mechanisms invisible.

In comparison with our other two examples, IoTA's engagement with hardware and public interest scientific knowledge is more oriented toward ends than means. The accessibility of hardware and electronics makes it possible to design civic data collection tools that use the same kinds of calibrations as the tools used by governments, but repositions the site of data collection so that communities whose interests may not be represented in official data collection can offer their data as part of their political voice. This constructs legitimacy in relation to constituted authority: the goal is to produce data that are valid on the terms that scientific and policy practitioners establish. The end goal of producing such valid data supersedes—to an extent—the means of participation that are the focus of other civic science projects.

Conclusion

DIY and hacking culture operate by undermining and appropriating systems and structures through material practice. This is more critique than integration, of institutional knowledge. The use of scientific hardware and measurement practices to collect and represent data coming from an alternative point of view illustrates some of the politics that can lie beneath engagements between adaptive and constituted authority. Producing, creating, curating, and contextualizing data obtained through scientific equipment or using scientific methods may provide an entry into broader political or policy discussions. This is a departure from many of the ways that hacking culture has been connected with scientific knowledge and the public interest.

The examples developed in this article illustrate how the development of legitimacy in relation to participation has often characterized the way hacking has engaged with institutionalized frameworks. Participation comes to be associated with forms of governance that are understood as valuable for market capitalism or even for the development of "collaborative dissemination" in science. There are advantages of this: an ethic of participatory knowledge creation as developed through the CERN OHL or a process of empowerment through appropriating science in a DIY ethic. But there are disadvantages too: that the development of the more radical outcome of accessible knowledge commons could be weakened by too much focus on "adaptive" authority and participatory governance, leading to expert rule and problems of gender-, class-based and racial exclusion.

In other words, the means of participation can limit the ends of shareable knowledge creation. Is the solution to try to engage with science and policy on the terms that their "constituted" authority establishes? What if this further mystifies science and technology, countering the efforts of DIY and hacking culture? As this article illustrates, hacking culture evokes as an end goal the accessibility of knowledge, but its valorization of participation can limit the achievement of these ends. This is entangled with the ways that legitimacy is understood within hacking culture and within the scientific cultures that open source projects now engage. Valorizing adaptive authority of participants, though it democratizes scientific knowledge and decenters some kinds of scientific expertise, still strengthens the focus on means, rather than the end goals of scientific investigation –unfortunately still often explained in terms of scientific legitimacy. The analysis here suggests that hacking culture has indeed made a difference in ideas about how to produce open knowledge, but that the outcomes of that production have not always produced the radical openness that hackers (and others) seek.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Barron A (2012) Kant, copyright and communicative freedom. *Law and Philosophy* 31(1): 1–48.

Barron A (2013) Free software production as critical social practice. *Economy and Society* 42(4): 597–625.

Benkler Y (2006) *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven, CT: Yale University Press.

Benkler Y (2011) *The Penguin and the Leviathan: How Cooperation Triumphs Over Self-Interest*. New York: Random House.

Berdou E (2011) *Organization in Open Source Communities: At the Crossroads of the Gift and Market Economies*. New York: Routledge.

Bodó B (2014) Hacktivism 1-2-3: how privacy enhancing technologies change the face of anonymous hacktivism. *Internet Policy Review* 3(4). DOI: 10.14763/2014.4.340.

Boisot M (2011) Generating knowledge in a connected world: the case of the ATLAS experiment at CERN. *Management Learning* 42(4): 447–457.

Boltanski L and Chiapello E (2005 [1999]). *The New Spirit of Capitalism* (trans. G Elliott). London: Verso.

Cammaerts B (2011) Disruptive sharing in a Digital Age: rejecting neoliberalism? *Continuum: Journal of Media and Cultural Studies* 25(1): 47–62.

Castoriadis C (1987 [1975]) *The Imaginary Institution of Society* (trans. K Blamely) Cambridge: Polity Press.

Coleman G (2012) *Coding Freedom: The Ethics and Aesthetics of Hacking*. Princeton, NJ: Princeton University Press.

Coleman EG and Golub A (2008) Hacker practice: moral genres and the cultural articulation of liberalism. *Anthropological Theory* 8(3): 255–277.

Collins HM (1998) The meaning of data: open and closed evidential cultures in the search for gravitational waves. *American Journal of Sociology* 104(2): 293–338.

Collins H (2010) *Tacit and Explicit Knowledge*. Chicago, IL: University of Chicago Press.

Collins HM and Evans R (2002) The third wave of science studies: studies of expertise and experience. *Social Studies of Science* 32(2): 235–296.

Dafermos G (2001) Management and virtual decentralised networks: the Linux project. *First Monday* 6. Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1481/1396 (accessed 20 February 2015).

Dafermos G (2012) Authority in peer production: the emergence of governance in the FreeBSD project. *Journal of Peer Production* 1: 1–11.

Dafermos G and Söderberg J (2009) The hacker movement as a continuation of labour struggle. *Capital & Class* 33: 53–73.

DiSalvo C (2014) The Growbot Garden Project as DIY speculation through design. In Ratto M and M Boler (eds) *DIY Citizenship: Critical Making and Social Media*. Cambridge, MA: MIT Press, pp. 237–247.

Dunbar-Hester C (2010) Beyond "Dudecore"? Challenging gendered and "Raced" technologies through media activism. *Journal of Broadcasting & Electronic Media* 54: 121–135.

Dunbar-Hester C (2014) Radical inclusion? Locating accountability in technical DIY. In: Ratto M and Boler (eds) *DIY Citizenship: Critical Making and Social Media*. Cambridge, MA: MIT Press, pp. 75–88.

Feenberg A (2008) *Questioning Technology*. London: Routledge.

Fluckiger F (2012) *Final Report of the Open Source Software Licence Task Force*. CERN. Available at: http://cds.cern.ch/record/1482206?ln=en (accessed 8 September 2014).

Fuster Morell M (2010) *Governance of online creation communities. Provision of infrastructure for the building of digital commons*. Unpublished doctoral thesis, European University Institute, Florence.

Jordan T (2008) *Hacking: Digital Media and Technological Determinism*. Cambridge: Polity.

Kelty CM (2005) Geeks, social imaginaries, and recursive publics. *Cultural Anthropology* 20(2): 185–214.

Kelty CM (2008) *Two Bits: The Cultural Significance of Free Software*. Durham, NC: Duke University Press.

Knorr Cetina K (1999) *Epistemic Cultures: How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Kostakis V (2010) Peer governance and Wikipedia: identifying and understanding the problems of Wikipedia's governance. *First Monday* 15(3). Available at: http://firstmonday.org/ojs/index.php/fm/article/view/2613 (accessed 20 February 2015).

Kostakis V, Niaros V and Gioitisas C (2015) Production and governance in hackerspaces: a manifestation of Commons-based peer production in the physical realm? *International Journal of Cultural Studies* 18: 555–573.

Krige J (2001) Distrust and discovery. The case of the Heavy Bosons at CERN. *Isis* 92(3): 517–540.

Lakhani KR and Panetta JA (2007) The principles of distributed innovation. *Innovations* 2: 97–103.

Lessig L (2006) *Code and Other Laws of Cyberspace (Version 2.0)*. New York: Basic Books.

Levy S (1984) *Hackers: Heroes of the Computer Revolution.* New York: Doubleday.

Mateos-Garcia J and Steinmueller WE (2008) The institutions of open source software: Examining the Debian community. *Information Economics and Policy* 20(4): 333–344.

Mansell R (2013) Employing digital crowdsourced information resources: managing the emerging information commons. *International Journal of the Commons* 7(2): 255–277.

Mansell R and Berdou E (2010) Political economy, the Internet and free/open source software development. In: Allen M, Hunsinger J and Klastrup L (eds) *International Handbook of Internet Research*. Boston, MA: Springer, pp. 341–361.

Mateos-Garcia J and Steinmueller E (2008) The institutions of open source software: examining the Debian community. *Information Economics and Policy* 20: 333–344.

Mateos-Garcia J and Steinmueller WE (2008) The institutions of open source software: Examining the Debian community. *Information Economics and Policy*. 20(4):333-44.

McQuillan D (2014) The countercultural potential of citizen science. *M/C Journal* 17(6). Available at: http://journal.media-culture.org.au/index.php/mcjournal/article/view/919 (accessed 23 February 2015).

Milberry K (2014) (Re)Making the Internet: free software and the social factory hack. In: Ratto M and Boler M (eds) *DIY Citizenship: Critical Making and Social Media*. Cambridge, MA: MIT Press, pp. 53–64.

Moody G (2002) *Rebel Code: Inside Linux and the Open Source Revolution*. New York: Basic Books.

Ostrom E (1990) *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Perens B (2014) Interview: ask Bruce Perens what you will. *Slashdot*. Available at: http://interviews.slashdot.org/story/14/04/03/1350216/interview-ask-bruce-perens-what-you-will (accessed 8 September 2014).

Powell A (2012) Democratizing production through open-source knowledge: from open software to open hardware. *Media, Culture & Society* 4: 691–708.

Powell A (2015) Open culture and innovation: integrating knowledge across boundaries. *Media, Culture & Society*. Epub ahead of print 5 February. DOI: 10.1177/0163443714567169.

PublicLab (2014) *PublicLab Annual Report 2013*. New York: PublicLab.

Ratto M (2011) Critical making: conceptual and material studies in technology and social life. *The Information Society* 27(4): 252–260.

Ratto M and Boler M (2014) *DIY Citizenship: Critical Making and Social Media*. Cambridge, MA: MIT Press.

Sauter M (2013) "LOIC Will Tear Us Apart": the impact of tool design and media portrayals in the success of activist DDOS attacks. *American Behavioral Scientist* 57(7): 983–1007.

Soderberg J (2008) *Hacking Capitalism: The Free and Open Source Software Movement*. London: Routledge.

Stallman R (1999) The GNU operating system and the open source revolution. In: DiBona C, Ockman S and Stone M (eds) *Open Sources- Voices from the Open Source Revolution*. London: O'Reilly and Associates, pp. 53–70.

Star SL (2010) This is not a boundary object: reflections on the origin of a concept. *Science, Technology, & Human Values* 35(5): 601–617.

Superflux (2014) IoTA: Internet-of-Things Academy, Phase 2. Available at: http://www.superflux.in/work/iota-phase2 (accessed 20 February 2015).

Tapscott D and Williams A (2006) *Wikinomics: How Mass Collaboration Changes Everything*. New York: Portfolio.

Thomas D (2002) *Hacker Culture*. Minneapolis, MN: University of Minnesota Press.

Toffler A (1970) *Future Shock*. New York: Bantam Books.

Von Hippel E (2005) *Democratizing Innovation*. Cambridge, MA: MIT Press.

Wark M (2004) *A Hacker Manifesto*. Cambridge, MA: Harvard University Press.

Wark M (2013) Considerations on a Hacker Manifesto. In: Scholz T (ed.) *Digital Labor: The Internet as Playground and Factory*. New York: Routledge, pp. 69–75.

Warren J and Regalado C (2014) Response to "How do DIY processes encourage citizen participation?" *PublicLab Research Notes*. Available at: http://publiclab.org/notes/Cindy_ExCites/06-25-2014/response-to-how-do-diy-processes-encourage-citizen-participation (accessed 20 February 2015).

Weber S (2004) *The Success of Open Source*. Cambridge, MA: Harvard University Press.

Wilbanks J (2007) Cyberinfrastructure for knowledge sharing. *CTWatch Quarterly* 3(3). Available at: http://www.ctwatch.org/quarterly/articles/2007/08/cyberinfrastructure-for-knowledge-sharing/ (accessed 20 February 2015).

Author biography

Alison Powell is Assistant Professor in Media and Communications at the London School of Economics and Programme Director of the MSc in Media and Communication (Data and Society). Her research examines how people's values influence the way technology is built, and how discourses, practices, and governance structures are produced in relation to new technological systems.

Data Activism

Alessandra Renzi and Ganaele Langlois

Introduction

Let us start with a commonplace observation: whoever owns, controls, and has the right to access and analyze data holds tremendous power over individuals and populations. This is true for governments that collect data on their citizens to develop policy and provide or eliminate social services and of social media corporations that gather, analyze, and sell all kinds of user data about consumer preferences and behaviours. And let us continue with a correlative statement: the power of data not only resides in its capacity to produce knowledge, but also in its ability to shape perceived realities. Data maps, graphs, and visualizations commonly circulate in both mass and social media to show us what our world is like, how we will be impacted by environmental or social changes, what kind of communities and individuals surround us, and whether these communities and individuals are friends or foes. Data shapes realities not only by enabling certain representations of the world around us, but also by enticing us to internalize these realities and make them our own. Data clearly has transformative and affective potential: the most powerful data visualizations are intuitive in that they immediately convince us that they make sense, that they are truthful, trustworthy, and empowering. They can, in turn, foster feelings of elation or fear, and they have the power to shape our sense of belonging to diverse communities.

Data, in short, yields tremendous *political* power and we rely more and more on data to understand and navigate the complexities of our individual and collective realities. It comes as no surprise that data therefore has an important role to play in civic life, and that activists are drawing on data as a way to provide means for social transformation. Political organizing and real-time communication through social media are some of the ways in which data is mobilized in activism. Many studies are understandably critical of the impact of social media on activism because of how social media tie autonomous social justice projects to dominant corporate players—i.e. YouTube, Twitter, Facebook, and the like—who collect and mine data and exploit free labour (Terranova 2000). Researchers have discussed this tension between autonomy and social control, as well as the risks social media pose to privacy and surveillance (Lovink and Rasch 2013). And yet, fewer have paid attention to how the sociality that emerges in this tension between freedom and capture reshapes activists' and other political actors' individuality and collectivity, redefining modes of solidarity, participation, and knowledge production along shifting notions of community, agency, and engagement.

Our aim in this chapter is to understand some of the new data-based activist practices and the ways in which they challenge and resist existing power relations. We want to hone in specifically on how activism engenders new modes of being and acting together through a direct engagement with data and the means of its mobilization. Thus, we look at the socio-technical field of seemingly cold or "objective" knowledge and facts in order to examine a live and fluid system of negotiations of individual affects, group belongings, and transformation of social bonds and that builds on data and organizing metadata.

Let us give an example of this with the recent case of protests sparked by the killing of unarmed African-American teenager Michael Brown by a white police officer in Ferguson, Missouri. As news of the killing circulated, people mobilized both on the streets and online to discuss and protest institutional racism in the United States. One way in which such mobilizations were reflected online was through the use of hashtags, a particular kind of metadata. On Twitter, hashtags like #handsupdontshoot and #Ferguson connected people across spaces and in some cases fostered unexpected solidarities: during the protests, Palestinians stood by the people of Ferguson and shared advice on how to deal with teargas and militarized police (Aljazeera 2014). The coupling of hashtags #Gaza and #Ferguson in the same tweets was not just symbolic. The ability to pool metadata enabled the circulation of information about the same teargas used in both places. It made a statement about the shared realities of oppressed peoples and created a bond among those affected. This capacity to impact individuals and groups is at the heart of understanding the power of data in the fostering of new activist practices.

Such instances of change cannot be understood as unfolding along the binary of technology, on one side, and collective practices, on the other. Rather, as will become evident when we trace the relations among disparate fields (social, cultural, and technological), processes (communication and action), and actors (activists, artists, and researchers), technology and collective practices are today indissolubly linked. Thus, in order to trace this transformative power of data while attending to both technological and social forces, we rely on the concept of *transindividuation* (Simondon 1989b; Stiegler 2013). Simply put, transindividuation designates the socio-technical context, or milieu, through which transformation unfolds, allowing for individuals to gain new awareness and to bond with groups that also evolve and mutate in reaction to events, other groups, and individuals. Transindividuation is an evolving process of co-constitution between the individual *I* and the collective *We*, and this process is often now produced, mediated, and transformed through data. In this sense, data activism is not separate from other forms of activism. The examples we examine in this chapter show how data activism is part of on-the-ground activism because of the sociality of the practices involved and because of the intimate relations that individuals develop through technological means in general, and data in particular.

This chapter examines some of these processes of individual and collective transformation that are mediated by data and that trigger the emergence of new activist practices. But first, a discussion of social transformation and transindividuation in relation to data is needed. This will be followed by three vignettes of different activist contexts where data mediates affective bonds (Occupy Streams and metadata), creates new forms of shared knowledge (Occupy Data and big data), and new vectors of transformation (Facial Weaponization Suite and biometric data). In the following discussion we argue that, in addition to studying the role of data for political and economic control, we need to pay attention to the different ways in which data is implicated in the circulation of mediated and unmediated psycho-physiological stimuli (affects, perceptions, and emotions) precisely because these stimuli generate self-perceptions, belonging, and collective action.

Transindividuation and Data Activism: Socio-technical and Political Considerations

Since the growth and availability of data, especially so-called big data, we have witnessed the emergence of a data paradigm that ties social dynamics to economic and political interests. Data is increasingly used to analyze and understand the behaviour of individuals and groups; the knowledge gained is employed to organize social life in all of its aspects, from the intimate (e.g. through targeted advertising on social media that piques desire) to the public (e.g. through policymaking and surveillance). In this context, we see data as socio-technical: it consists of technologically produced sets of informational resources that are mobilized within social, economic, and political processes. Data activism takes place at the crossroads of the technological and communicational logics feeding capitalism. It attempts to wrestle the socio-technical power of data from the hands of dominant groups to promote social and economic justice.

In order to understand how data mobilizes and is mobilized in the context of activism, it is necessary to first discuss the relationship between individual and collective action and data. Often, to explain social change we posit the pre-existence of at least one of these two reference points—the individual and the collectivity—as already-formed and distinct entities partaking in collective action through rational choices. We often hear about how actors such as activists *act* within collectivities to promote social change: for instance, groups of citizens dissatisfied with the government response to the 2008 financial crisis decide to occupy public space giving rise to the Occupy movement in many parts of the world. In other cases, someone may mobilize members of their community to launch an advocacy campaign, as was the case with the online petition that called for the prosecution of George Zimmerman, the man released without charges after killing unarmed 17-year-old African-American Trayvon Martin. The petition sparked a series of anti-racism protests and “hoodies walks” all over the United States. In both examples it is easy to focus on the preexistence of individuals and collectivities that are mobilized around specific issues; but this glosses over the complexity of how collective action comes to be.

Drawing on the concept of transindividuation, we can look at the relationships that *generate* both individual and collective action and conceive of the individual and collective as the result of a socio-technical genesis. In the case of the Trayvon Martin campaign, it would be reductive to simply claim that the individual who initiated the petition sparked collective action. Rather, specific social and data-related events had to take place before the campaign could gain public attention and mobilize supporters. In fact, as David Karpf explains elsewhere in this volume, members of the online petition company Change.org came across a petition circulating on a mailing list by someone who had read about the killing in the news. Change.org staff decided that the issue had the potential to go viral and contacted Martin's family to start a new petition with them. High profile sponsors, algorithmic calculations, and a professional social media campaign heavily relying on metadata guaranteed that the call to action reached large audiences (and not without financial gain for Change.org). Here, the context in which transindividuation engenders individual supporters, as well as the marching crowds that hit the streets, can be described as one wherein a variety of factors, including data analysis and complex algorithms that extract surplus value from social justice causes (while supporting them), viral information circulation, and the emotions circulating on social media networks converge to establish a new relationship between the one and the many. This is how transindividuation takes place within a socio-technical field where the two poles of individuality and collectivity emerge simultaneously as the *result* of newly established relations (instead of having individuals and collectivities be the preexisting terms of a relation). Importantly, transindividuation is not simply a technological process but one wherein technology, and in our case data, is a vector for the circulation of affective and emotional bonds. If we think of transindividuation as layered, the genesis of individuals and collectivities unfolds at multiple levels, from the micro-sensorial (affects and emotions) to that of action—both individual and collective action. And as the previous Ferguson example demonstrates, data bridges individuals, modulating the relation between the *I* and the *We*—our sense of ourselves both alone and as members of a community.

To illustrate how affective modulation takes place, we can think of how people come together around a protest. There are different ways of participating in protests: as organizers, as affinity groups (e.g. a pink or black bloc); there are those who walk alongside the march "lending bodies," bystanders who cheer or honk, and those who follow through media. These different participants will share an open field of intensive relations where affects ultimately connect them. According to Simondon, both individuals and collectives are fluid entities that are always in a metastable-equilibrium, i.e. an intensive state where change is triggered by external events that alter their existing equilibrium. These changes can be as small as a sensation or as far reaching as a crisis—they trigger flows of transformation. Indeed, as a context for transindividuation, a protest is the site for the circulation of mediated and unmediated psycho-physiological stimuli that are constantly reorganized into feelings and emotions along a personal-social continuum.

In this context, sensorial and bodily experience—so-called affects—precede and trigger emotions, which then prompt action. It is through affect and emotions that the tension between constituted individuality and the collective is first felt; affect regulates the relation between an individualized being and the pre-individual milieu (the "interiority" of an individual) as it encounters the world; emotion, arising from the difficulty of rendering an affective plurality into a unitary meaning, engenders the collective when structured across many subjects (Simondon 2006, 111–22). One may be motivated to join in a protest, or chant, or confront the police because of feelings that emerge through these relations. Indignation, for example, is a feeling so wide-spread lately that it has even become the name of an entire social movement in Spain (the *Indignados*). Data often mediates and modulates such processes. Given the intense use of social media during protests, data plays a key role in the mediation of affects that modulate the transindividual field of relations—what Simondon calls the transindividual milieu (Simondon 2006). For instance, the metadata of tweets choreograph protests by moving crowds as they communicate about events, routes, and encounters with the police and media (Gerbaudo 2012). In this sense, data is a vector of affects that itself has a kind of agency.

VersuS' real-time visualization of a protest in Rome on October 15, 2011 provides an animated map of the affects and emotions circulating at different times of the day, with intensity literally peaking as police attack protestors and more people join in the streets (Art is Open Source 2011). The map uses data collected from major social networks like Facebook, Twitter, YouTube, Instagram, Foursquare, and Flickr, and analyzes it through natural language analysis and artificial intelligence that isolates words indicating affects, emotions, and participation. The visualization itself only makes evident the extent of the engagement with social media during this attempt at social change. Still, it helps us consider how the socio-technical space of social media networks and the proximity of bodies with mobile technologies on the streets of Rome function together to create "milieus" where circulating feelings like joy and belonging or outrage and panic reposition individuals within and among groups. We understand this field as a transindividual milieu composed of the affects and emotions circulating in connection with specific technologies (what Simondon calls "technical individuals"), technical ensembles (the discourses and contexts that produce and make sense of technology), individuals, and the collective. In the Rome protest, both individuals and collectivities are reshaped through the mediated process of communicating about and actually engaging in protest in the streets. These experiences are processed differently in each individual because of the unique interiority of the individual processing the stimuli—their "pre-individuality" to use Simondon's term. Nevertheless the emerging relations bring people to act collectively along shared intensities. In this context, the VersuS map can be read as a real-time simulation (rather than visualization) of the intensities that traverse the transindividual milieu and affect those that inhabit it. It is a simulation of the process of *becoming collective as activity*, i.e. as a set of practices, perceptions, significations, communications, and so on; it poses the problem of intersubjective relations not at the level of discourse but at the level of affectivity and emotivity that are deeply intertwined with the data enabling and channelling their expression (Simondon 1989, 13).

The implications for thought of the transindividuated subject are very different from those of the preformed individual belonging to an existing community. Focusing on the latter subject, we argue, is not helpful for answering questions about the role of data in the feasibility, scalability, and durability of activist formations because it does not interrogate how movements emerge. Rather, by refusing the individual as the measuring unit of the collective, we can pose the problem of the collective from the perspective of its members as sets of affective, communicative, and embodied relations that are mediated by data, and where the process of becoming can be investigated. What interests us is not how groups retain their identity or structure but how they change (Simondon 2006) and how they do so in relation to data. The concept of transindividuation enables us to look at data in the form of information, metadata, and algorithms as structuring elements of a transindividual milieu wherein the individual and the collective emerge simultaneously through relations shaped by circulating affects. Contemporary forms of data activism act within the transindividual milieu to deconstruct, create, and realign specific articulations of the social and the technical in order to enact change. And, as we will see, these transformations also take place in different kinds of socio-technical environments that are not necessarily connected to social media, like those for the collaborative production of knowledge through big data, or the creative use of data for radical community art projects.

To further frame our case studies, we are guided by Felix Guattari's remark about the post-media era (2012). While Guattari did not live to witness the rise of contemporary forms of digital technologies, he nevertheless saw emerging digital technologies of the 1990s as ushering in a post-mass media era, which he thought would be composed of decentralized networks, affiliations, and associations that would act much in the same way as localized activist movements engage in broader global alliances. New technologies of communication, Guattari argued, could be appropriated in order to deconstruct power formations, to give birth to new creative processes and thus foster new subjectivities and new ways of being and living together in the world. The role of communication technologies was to be the vector through which new creative and resistant relationships could be transmitted. What mattered was not only the appropriation of the media themselves, but the mobilization of alternative media as sites of experimentation with new social relationships. The idea was that communication technologies would increase the possibilities of creating radically new ways of being together that would escape and neutralize dominant and unequal economic and social relations. In Guattari, we already find some key modes of activist transindividuation in his experiments with pirate radio stations in France and Italy, where he was a contributor and supporter. The pirate radios were not just producers and distributers of alternative information. Their experiments with the form and content of radio transmission (experimental shows, phone-in contributions, poetry readings, street parties, etc.) fostered encounters among a variety of groups—the activist and community groups contributing to the programming and running of the project. These encounters engendered decentralized alliances and ways of being together that crafted different social imaginaries.[1]

Much work has already been done examining how Guattari's post-media era is reflected in contemporary forms of activisms, although not so much for data activism. Software activists have been working to create alternative social media platforms allowing for

[1] Guattari was interested in the kinds of multiple refractions that new techno-social alliances could produce. What he called "co-individuation" among groups and individuals, in other words, could not be understood using linear causality models: it required understanding both macro and micro practices and states of being, from one's psychic experience to broad economic systems. In the same way, transindividuation cannot be understood from a linear perspective: rather, transindividuation involves echoes, resonances, and refractions that spread in no logical fashion. As such, for Guattari, the work of activism was not simply one of targeting a specific area, such as the economy or the environment, but rather of working at the intersection of different ecologies (Guattari 2000): political, social, and economic, of course, but also communicative and subjective through the refashioning of human bonds and modes of collective imagination.

decentralized alliances and protecting users from surveillance through greater anonymity. The greater availability of mobile media recording devices along with fast access to the Internet has provided for alternative forms of expression and new modes of sharing information (e.g. Indymedia). The crafting of social imaginaries along with new subjectivities emerging from new ways of being together has been at the centre of contemporary activist practices. One could think of the *Indignados* from Spain, where the sentiment of outrage at current inequalities and economic policies fuels alternative social, economic, and political responses. What matters there is the creation of new perceptions through the deconstruction of power formations, for instance, the dissociation of higher education from the imperative of economic return on investment. In the same way, the Occupy movement was both a denunciation of current inequalities—the famous "We are the 99%" slogan—and an experiment in new ways of living together through, for instance, general assemblies, consensus-based decision-making and human megaphones. Crucially, the concept of transindividuation along with Guattari's theories on the post-media era, open up a discussion about the socio-technical character of the new alliances and practices that emerge, or may be intentionally fostered among activists. The following three vignettes take up this task: they map the composition of new socio-technical activist formations, identifying the kinds of data that engender new relations, exploring how affect and emotions circulate in the transindividual milieu, and describing the flows of individual and collective transformation that mobilize actors.

Occupy Streams: Data and Social Media Platform Politics

Our first case study is about the connection of activism and data through social media platforms. Occupy Streams, along with similar platforms, helps us better understand activist practices based on changing notions of solidarity and participation. Such notions, in turn, are shaped through the tension between autonomous and commercial platforms, which channel users' affective and emotional reactions to circulating information about political events. On these platforms, data takes the form of content—textual and increasingly visual, but also metadata like hashtags—and circulates through information objects such as "like" buttons. We argue that it is important to investigate how these different manifestations of data come together with human agents into a shared socio-technical field where new forms of participation and solidarity emerge.

In general, social media platforms allow for the collection, storing, and distribution of digitized content, from video to comments. Social media platforms are meant to manage this digitized content by classifying and organizing it. This requires the use of metadata, that is, data that provide a piece of information about the contents and context of other data (e.g. the time a video was uploaded). Since metadata facilitates the search and circulation of data, it helps activists reach wider audiences. It also helps to make connections across issues and to circulate information across social and technical environments. As already mentioned, the use of metadata such as hashtags pools content together, enabling the spontaneous emergence of counterpublics around specific events or issues (Warner 2002). Finally, metadata also enables new kinds of archiving and preservation practices that contribute to the creation of an embedded social movements memory. Thus, paying attention to the socio-technical composition of this field where activist practices play out demands that we recognize the ways in which metadata itself has a certain kind of agency. Metadata corrals activism into processes of capital accumulation (e.g. through data mining) but it can also create the conditions for the development of new activist practices (e.g. campaigns around hashtags). Most importantly, it is actively implicated in circulating affects and fostering social relations.

Since its beginnings in the 90s, autonomous media in general, and Independent Media Centres (IMC or Indymedia) in particular, have changed a lot. In addition to mobile technologies that have made the necessary infrastructure (internet connections, video cameras, etc.) increasingly accessible, open publishing platforms are more sophisticated, allowing for immediate posting of text, images, and video and hosted discussions. New media platforms can integrate the traditional features from Indymedia with feeds from various social media and live streaming. For example, the platform used during the anti-G20 convergences in Pittsburgh and Toronto, and during the protests against the Vancouver Olympics, automatically published YouTube videos, tweets, text, and a map to locate events as they were happening. Livestreaming, in particular, has become very popular during the wave of protests that followed the financial crisis in Europe and North America. Commercial platforms Livestream and Ustream provide the infrastructure to sites like Global Revolution and Occupy Streams to output and centralize 24-hour coverage of protest camps around the world, while chats and Twitter and Facebook feeds offer an interactive experience for activists and viewers (Costanza-Chock 2012; Juris 2012).

At the formal level, streaming technologies have changed the content as well as practices of radical media from reports and analysis to embedded journalism and live correspondence. The availability and transferability of standardized platforms and media activist toolkits has created a sort of "media centre franchising," where different citizen reporters' websites offer similar interfaces and features like social media buttons, embedded chats channels and live-casts. On these platforms—and across them—the metadata behind the familiar features organizes content in a way that is intelligible to the majority of users. Much of the similarity and uniformity of the interfaces comes from incorporating black-boxed commercial social media modules like Twitter feeds or share buttons into activist projects' websites. Not unlike the franchising of a successful business model or brand, corporate-built, customizable platforms or modules are made available to users who will customize them for their goals. Instead of paying a fee, the activists will make the collected metadata available to the different companies embedding their services in the platform.[2] To add to this franchising analogy: familiarity with the services and brand (for instance, different Occupy streaming channels) more easily engage users, plugging new actors into networks where they feel at ease with the interfaces and comfortably take on the roles of reporters, participants, moderators, and voyeurs at protests and other political events. Here, the habituated gestures of engaging with commercial platforms for social networking—the sharing and endorsing for

[2] This happens by signing the Terms of Service for the use of a platform.

instance—are embedded directly into political practices simply as a result of their new context.

Thus, despite their problematic ties to corporate actors, new media activist toolkits have expanded the context for transindividuation: in the constant flow of information at rallies and encampments, the unleashing of affects and emotions not only impacts public perception of actions by denouncing violence and portraying experiments in direct democracy, it also redefines participation, allegiance, and group boundaries through the recovery or invention of political and cultural practices. This happens as individuals connect to others through circulating data. For example, reporting within social movements, today, seems to be done more by single individuals than by groups and it tends to connect people who identify as movement reporters on shared platforms, rather than in the physical space of a media centre. These mediated relations among activists are an important indicator of the changes to the composition of social formations, which stem from specific social conditions (e.g. a stronger sense of individualism that is promoted by social media) and new political needs (e.g. to break through overcrowded social media channels). At the same time, the changes fold in established traditions of struggle (e.g. media activism) and may build on practices of cooption that are latent or present in hidden forms at sites of conflict (e.g. the repurposing of commercial platforms).

Looking at activist social media data and metadata and their function as producers of affective and emotional relations (for instance, relations of solidarity) and as vectors of agency, we see these practices of resistance as not limited to the *communicative* action often ascribed to social media public spheres. Rather, they exist alongside and in connection with *direct* action. For instance, the live feed of CUTV during the 2012 Quebec student protests was a fundamental tool to update and draw to the streets students who followed this established university channel and became involved in five months of intense mobilizations against tuition hikes. Studying the deployment and function of CUTV and the circulation of its data offers a map of the emerging affective relations between activists, students, and wider publics that were key in sustaining the protests at a high intensity for an extended period of time. Similarly, the streaming of the Occupy Toronto channel functioned as a monitoring system to quickly mobilize critical mass at the encampment to prevent impending evictions. In addition to being part of phone trees, following this encampment's stream was an activist tactic to guarantee a quick response in case of emergencies, rather than simply a way to keep up with current affairs.

In these examples, the socio-technical milieu for transindividuation that emerges in connection with new manifestations of media activism is composed of human, machines (what Simondon calls technical individuals), technical elements, resources, and *data*, both as metadata and as information in its more traditional sense of content. As part of wider socio-technical milieus, social media platforms themselves consist of a variety of interconnected technical elements that resonate with each other, modulate our relations to technology, and mediate our interaction with others. These streaming platforms extend and connect life at protests and camps to the outside and to other platforms; they are sites where technology plays a vital role because of how we develop relationships with it, and to others through it. Networked social media, both autonomous and corporate, are not only the means or tools to connect individuals but they are active agents that carry with them an associated milieu where the functions of technology are organized, reproduced, and in some cases contested. This aspect too speaks to the agency of technological elements in the socio-technical milieu.

In fact, we might think of platforms themselves as "technical individuals," modern machines that bridge humans and the natural word (Simondon 2006, 262) and also as the site where the machine and its associated milieu simultaneously emerge in a pattern of recurrent causality (Simondon 1989a). The "organs" of the platform—its interface, buttons, chat boxes, algorithms—connect it with other platforms into ensembles that enable reproduction of the platforms' functions but also their transformation. Here, we can think of the ways in which algorithms and buttons, as well as log-in functions, have enabled the development of interconnected platforms as a way to extract value from data. Yet these processes are not only economic but also affective. Since the use of platform's features is by now almost "intuitive," standardized platform elements like the chat boxes, Twitter roll, and related live channels framing the main feed of sites like Occupy Streams can be thought of as highly concretized forces that function in a variety of milieus and can work together in various kinds of machines and technical ensembles (Lamarre in Combes 2013, 104–05). As mentioned, the interface features that travel from commercial technical individuals like YouTube to radical media ones harness feelings of familiarity, participation, and interactivity that tap into a discursive field emerging from and engendering social media as specific technical individuals (this is what Simondon calls the technical ensemble).

This means that standardized streaming platforms, which seemingly leave very little space to do other than create a voyeuristic experience, can yield insights into the continuum between communication and action, individuality and collectivity, because of the way they enter into composition with specific activist formations. The individual's engagement with a platform's different elements—the interaction with the algorithms linking and filtering content, and then creating patterns of meaning from this—is rife with moments of intensive affectivity and emotivity that are dependent on presence (encounters) and action. As mentioned earlier, the individual does not pre-exist the collective, and since it carries with it a shared field of metastability that makes it always open to change, we can conceive of the collective as a system of relations that is itself transindividual (Combes 2013, 40–43).

Many of the protest live feeds on Occupy Streams no longer broadcast and yet endure in an online afterlife in what resembles a haunted TV studio. Here, the frozen interface is a testimony to the processes of production that took place but also signals the potential for new relations. In particular, the metadata and code of the platform—its information objects—can engender new kinds of relationality that produce transindividuation. The interfaces leave different kinds of archives than previous forms of autonomous media. As opposed to the articles, images, and video of selected events that we found on alternative media sites up until five years ago, the archived streaming of activities, assemblies, and confrontation with police offers raw material for analysis and reflection, not just for scholars but *especially* for activists. The activist art installation *Printemps CUBEcois* by

David Widgington is an example of creative use of archival material to remember and reflect on the student protests in Quebec. This public art project installed in the atrium of Concordia University Student Union consists of a large cubic room covered in over 3,000 posters, banners, signs, and stencils collected from various social movements during and after the protests. Inside, TV sets broadcast CUTV footage of the rallies. Widgington's intention as an artist/activist is to activate a non-nostalgic archive for re-aggregation and collective self-reflection (Widgington 2013). The *CUBEcois* seeks to nurture the oppositional consciousness that was strong during the protests, reacquainting activists with past protest performance, in order to prepare for future struggles (Widgington 2013). The archive/installation functions as a site where affect and emotions "re-play" in relation to memory, reorganising the relation between the *I* and the *We* of the students and relocating individuality and collectivity within a shared history (rather than a shared moment as is the case at protests). This memory is not just looking into the past but is mobilized to situate the students in relation to a potential shared future. As relational artefacts, the banners in connection with the replayed video streams reconnect to the past but also produce anticipation because they can be used again in the streets. In this context, thinking about potential processes of transindividuation alongside Guattari's conceptualization of the post-media era opens a discussion about the potential for those in struggle to re-imagine the power of radical media platforms through different practices that take up new forms of archiving, meaning-making and analysis of the events as they happen and in their aftermath.

Occupy Data: Data Visualization and Knowledge Production

The previous case study focused on data as transformed media material, and on how the platform as a site for data management catalyzes new relations that build on the affectivity and emotions involved in witnessing, producing or remembering political events. This case study engages with big data as a source of new knowledge. Using the example of Occupy Data we discuss how data activism takes the form of a direct engagement with big data through data correlation and the production of new insights into the past, present and future. In this context, the process of transindividuation unfolds as data analytics are used to redefine the parameters of the possible through the creation of shared meanings, memories, and expectations. The knowledge yielded by big data (defined here as large sets of facts about an object) is commonly believed to offer a more comprehensive picture of the state of a situation and its future development. In the process of wrestling data from the hands of the dominant groups who control this picture, Occupy Data activists forge new alliances and push the boundaries of the imagination aided by independent data analytics. As a result, new relations are engendered that allow for the organic construction of individual and collective expectations that are not imposed through top-down mythologies.

Bernard Stiegler offers two useful concepts to understand how knowledge resulting from data modulates transindividuality: *retention*—what we retain from the past; and *protention*—what we come to expect of the future (Stiegler et al. 2011). In so doing, Stiegler usefully shows that individual and collective past, present, and future are constructed through what affects us, what is remembered and what comes to be expected. It is within these parameters that decisions are made, be they personal, collective, political, economic, social, and so on. In the big data paradigm, data is increasingly used to manage the consequences of neoliberalism. An example that comes to mind is data mining to extrapolate the consequences of global warming and climate change if no political action is taken. As Wendy Chun (2011) astutely argues, such visualizations serve to reify the future: they convince us that only one kind of reality is possible. Since this goes against the idea of the future as open to possibilities, we have a tension in the data management paradigm. The new knowledge that big data yield always runs the risk of narrowing a horizon of imagination about what could be, and thus further reinforce existing power structures. What is at stake in Occupy Data, then, is the management of the knowledge that serves to create memories and expectations that contribute to shaping self-perception and visions of the world.

Our sites of analysis—Occupy Data and the local Occupy Data New York—work within that tension, engaging in data collection, storage, and analysis to push the boundaries of what can be known or imagined. These Occupy Data initiatives directly address the question of the production, distribution, and ownership of data. As its name indicates, Occupy Data is an offshoot of the Occupy movement, and focuses particularly on strengthening "initiatives of the Occupy Wall Street Movement through data gathering, analysis, and visualization" (Occupy Data 2012). At the data production level, Occupy data works to make existing public data sets available to the general public. It also hosts projects that enable users to set up data collection. For instance, the "Data Anywhere Project" from February 2013 states that: "Data is available in bits publicly, but aggregated by companies that want to charge for it. Other data may be free in aggregate form, but not available for live query/access. This project aims to solve both problems, one data set at a time" (Occupy Data 2013). Occupy Data also makes data sets available to the public, along with tools and tutorials for data analysis. The site features reports from specific projects, some focused on the effects of the financial crisis (e.g. foreclosure statistics) and environmental disasters (e.g. Hurricane Sandy), others focusing on using data to understand the reach and dynamics of the Occupy movement. Last but not least, Occupy Data initiatives provide tutorials as well as organize events and hackathons where people can gather and work on common projects.

As such, Occupy Data engages with all levels of data management and highlights the different kinds of struggles that operate in the big data paradigm:

- Data collection: who has the right to collect data and how? Who owns data?
- Data storage and retrieval: who has the right to access data and under which conditions?
- Data analysis: what are the algorithms used to analyze data, and what kind of logics (commercial, social, for instance) is embedded in them? How does one define what data stands for?
- Data visualization and distribution: how is data represented as human-comprehendible information? How is processed data made accessible?

Each of these stages of data production and circulation presents us with specific forms of struggle around commercialization, secrecy, and ownership of data. As the struggles unfold, individuals come together in a variety of interactions that affect and change them.

Occupy Data argues for openness at all stages of data management by making both data sets and tools available. While corporate research argues for the right not to let participants know that they are subjects of research—e.g. the recent Facebook emotional manipulation experiment (Kramer et al. 2014) where a large number of users unknowingly had their news feeds manipulated to study mood changes—projects like Occupy Data urgently ask about the goal of such data research. They promote independent data gathering, transparency of research processes, and active participation in designing research projects and their parameters. Here, the collective visioning of reality does not only manifest itself in the final data analysis and visualization, but also and especially while responding to the challenges that this type of data activism faces—for instance, developing a research paradigm that would be different from the neoliberal one or discussing the ethics of data collection, storage, and retrieval.

The data visualizations produced look like many other kinds of mainstream information visualization. Occupy Data mobilizes standard open-source tools for analyzing large data sets through different kinds of semantic and geo-locative visualizations. They favour user-friendliness and readability by organizing and categorizing information in spatial and visual terms: semantic closeness becomes spatial clustering, and thematic links are rendered as actual lines. Data here engages participants and viewers in a redefinition of the past, present, and future, whereby the circulation of affects is structured across a collective that often comes to be a coherent body in the moment it assembles, or in the resulting visualization itself. Facts become visible by being organized as coherent, shared representations that are almost imperceptibly mediated by software for data analysis. Those producing or relating to the representations are able to make sense of themselves and their peers anew through a direct connection that is established and modulated by the technical interface. The parameters of belonging to this collectivity and the boundaries of the groups that compose it are reshaped in the resulting field of visibility: Occupy Data as a symbolic gesture that occupies dominant knowledge production, as well as the actual projects that visualize specific issues. Nowhere is this more apparent than in the case of data analysis and visualization of projects focused on the Occupy movement itself. The Occupy Data NYC page lists that 13 out of their 23 projects are focused on the Occupy movement itself. The projects mostly include visualizations of different aspects of the movements, from mapping police violence during the protests to visualising the themes and thematic links among participants and groups in the movement. Working from the inside, such visualizations tend to give coherence to a movement that has often been criticized for being decentralized, non-hierarchical, and disparate in terms of political and social demands.

Studying Occupy Data from the perspective of transindividuation draws attention to the uneasy link between activism and the established practices of data management. In particular, we want to briefly consider the issue of how activists using big data perceive the relationships between data and so-called reality. Data management is overall a positivist paradigm: it assumes that by extracting and analyzing facts about objects, a more precise picture of reality can emerge. As Bruno Latour (1993) and other critics of the positivist paradigm in techno-social research would argue, this is a fallacy: data management is about the construction of a specific reality, not the discovery of a pre-existing one. That is, the data management paradigm is not an objective measure of a world out there, but a specific construction of it. We have already talked about how the neoliberal paradigm used data management as a way to reify a specific conception of the world and its future development. But the issue is not only about questioning neoliberal assumptions built into the data research process, it is also about questioning how data management itself already offers some kind of reified representation of the world. The problem lies in the claim that such representations are pictures of reality. As Galloway (2012, 80) argues:

> Data, reduced to their purest form of mathematical values, exist first and foremost as number, and, as number, data's primary mode of existence is not a visual one. Thus (…) any visualization of data requires a contingent leap from the mode of the mathematical to the mode of the visual. This (…) means that any visualization of data must invent an artificial set of translation rules that convert abstract number to semiotic sign. Hence (…) any data visualization is first and foremost a visualization of the conversion rules themselves, and only secondarily a visualization of the raw data. (…) And because of this, any data visualization will be first and foremost a theater for the logic of necessity that has been superimposed on the vast sea of contingent relations.

It would be thus a mistake to think that data visualization automatically leads to a faithful representation of reality and that it is that representation which, in turn, enables processes of transindividuation. Rather, data visualization imposes a specific logic of correlation and relation among data points and individuals that creates the conditions for processes of transindividuation that build on events and encounters, but also on the knowledge and discourses that feed the processes of meaning making. In this sense, it is more useful to see data visualization as an entry point into an emergent field of socio-technical relations composed of temporary and biased representations that foster new alliances and forms of intersubjectivity. Here, studying the practices that reimagine the potential relations to data itself makes new modes of resistance to the data paradigm visible. These modes of resistance openly contest the politics of data-knowledge and adopt a more experiential and playful approach to envisioning and visualizing possible realities. The projects discussed in the final case study do so, using creativity to critique and subvert the dominant data paradigm.

Creative Data Activism

Activists and artists are indeed interrogating the politics of data knowledge and doing so in surprising ways. One such example is Zach Blas' Facial Weaponization Suite, which addresses the paranoia about the inability to recognize faces in a society where face recognition technology is increasingly prevalent. The project consists of a series of community-based workshops where participants make “collective masks” that are modeled from the aggregated facial data of participants, resulting in amorphous masks that do not register as human faces to biometric technologies. The masks produced are then used for public interventions and performances. The project troubles common assumptions about the objectivity of data and, in particular, of biometric data for facial recognition. The latter is unmasked as building on centuries of hetero-normative, racist, ableist, and classist (Magnet 2011) principles that set the standard of normalcy, against which alterity is constructed and monitored. Unmasking takes place by creating what Blas calls “social opacity” (2011). Masks like the Fag Face Mask*,* created by merging facial data of queer people, question the assumptions and consequences of scientific studies determining sexual orientation through facial recognition techniques. Another, the Black Mask, “explores a tripartite conception of blackness, divided between biometric racism (the inability of biometric technologies to detect dark skin), the favouring of black in militant aesthetics, and black as that which informatically obfuscates” (Blas 2011). Other masks focus on issues of visibility and concealment in feminism, and on migration, xenophobia, and nationalism. As Blas puts it: “These masks intersect with social movements’ use of masking as an opaque tool of collective transformation that refuses dominant forms of political representation” (Blas 2011). “Informatic opacity” is one way of resisting the positivist assumptions of data objectivity (Blas 2013). This approach brings data activism in alignment with other acts of escape and opacity that have increasingly marked new cycles of struggle to resist capture and recognition, from Anonymous and black blocs to Pussy Riot and the Zapatistas. Finally, the Facial Weaponization project also gestures towards the development of tech tools for encryption, anonymity, and privacy (Blas 2013).

While Blas’ project is described as one that creates “opacity,” we can, of course, see data activism like this as also rendering the power relations connected to data less opaque. As a form of data activism, the Facial Weaponization Suite, like many other collaborative design projects, points in the direction of what could be called “co-research-creation.” The important aspect of these projects is coproduction as a form of organization, one that is critical enough to know how to harness the technology within sociality or, when necessary, reject it to reclaim our social time and energies. Because of the co-involvement of activists and artists—the two often coexist, even in the same person—the production of knowledge and artefacts mobilizes politicized creative practices—what Guattari would call ethico-aesthetic practices (1995)—that are affective in the very way they foster political organization. Individual and collective transformation takes place first in the encounter and exchange between individual and data, and then as forms of oppression and forms of resistance come to light at the intersection of sociality and technology.

It is not only the masks themselves, and their exhibition out in the world, but also the workshop-based production of these masks that we consider a form of data activism. In fact, the critical discussions and the collaborative process behind the design and production of the different masks are important moments for a radical transformation of the participants’ relationship to media representations (Guattari 2000). In the collaborative process of producing the collective mask, we can see the unfolding of processes of transindividuation, first, in the sharing of the experience and engagement with the object, and then, in the actual mask. The mask symbolically and materially cuts the duality between the individual and the group by scrambling and reassembling the data that might otherwise be used to control them. Furthermore, the performances that incorporate the mask establish a relation to wider audiences than the workshop participants; they not only draw out knowledge about biometric data, and how it is collected and used, but also illustrate how transparency and structural discrimination affect the communities that develop the performances. The kind of community emerging as people interact with the mask (or masks) is the *product* of these interactions, rather than a predefined category to describe those involved in the project.

This kind of project resists data as a form of control. As opposed to the more traditional approach toward data management pursuing objective certainty, the Facial Weaponization Suite humorously serves to interrogate and trouble the common data paradigm. What is particularly interesting about the masks is that they show that data can be used to explore potentials rather than construct a specific version of reality. This kind of experimentation into what could be is playful, but in a serious way. The concept of serious play as an important step for activism is about engaging in experimentation with the world, where common assumptions are deconstructed and alternatives tested. In psychotherapy, this process is referred to as transitionality (Winnicott 1953): the space of experimental engagement with the world in order to reposition the individual in relation to others. In that regard, Blas' masks show how the distortion and refashioning of data can be used to question internalized frames of reference and open up new alternatives and new ways of thinking and being together.

Conclusions

This chapter initiated a discussion of the role of data in processes of transindividuation, wherein activists are mobilized as individuals and as members of collectivities. After discussing the relationship between activism, data, and transindividuation, we provided three examples of what we call data activism. The example of Occupy Streams discussed how new media activist practices relying on the circulation of data and metadata sustain and perpetuate affective bonds among those involved in protests. Occupy Data showed how big data analytics create a new, shared vision of the past, present, and future for activists involved in rethinking the role of data-derived knowledge, and in using data to contest dominant visions of reality. Finally, the Facial Weaponization Suite is an example of the creative use of biometric data as a vector of transformation that repositions profiled groups in relation to generalized fears and dominant discourses about transparency. These examples share an indirect and direct critical engagement with the data management paradigm that has become more common in the security and austerity cultures after 9-11 and the 2008 financial crisis: due to a widespread loss of faith in the securitized, free-market paradigm, citizens are increasingly questioning the rationale used to justify neoliberal discourses and policies.

Moreover, the projects discussed illustrate three different cases of the co-emergence of the *I* and the *We*. They emphasize how data facilitates the genesis of these two terms. Here, individuals and collectivities do not predate the relation that data affords; as entities that are in a meta-stable equilibrium, subject to change, they are engendered or constantly affected by such relations. In this context, projects like Occupy Streams, Occupy Data, and the Facial Weaponization Suite call for new research methods that make sense of the potential for change in a socio-technical field where data has a growing presence. We argued that to grasp this potential it is not enough to study the implication of data in mechanisms of political and economic control, whereby data is used to collect information about citizens/users and to organize civic life and consumption habits. We also need to pay attention to the different ways in which data is implicated in the circulation of mediated and unmediated psycho-physiological stimuli. These stimuli and affects, constantly reorganize the personal-social continuum of feelings and emotions and impact self-perception, belonging, and collective action.

Finally, we emphasized how positivist notions about data analysis' objectivity and reliability guide the study and use of data, especially big data (which sets the policy agendas in neoliberal economies). The modes of data activism we discussed reject this paradigm, directly addressing the question of the production, distribution, and ownership of data and questioning any claims about data's truthful representation of so-called reality. Therefore, research on data and activism should be adapted to the shifting reality of the challenges that the use of data poses for groups seeking to actively effect change. It is particularly important to embrace the complexity and paradoxes that are ignored by positivist approaches to the meaning and function of data. Ultimately, with this initial discussion, we would like to call for future research on data and activism that pays attention to the multilayered ways in which data in all its forms both fosters and restricts new socio-technical compositions.

References

Aljazeera. 2014. "Palestinians share tear gas advice with Ferguson protesters." *The Stream* August 14, 2014. Accessed September 30, 2014 http://stream.aljazeera.com/story/201408141902-0024060.

Alquati, Romano. 1975. *Sulla Fiat e altri scritti*. Milano: Feltrinelli.

Alquati, Romano 2003. "Intervista a Romano Alquati – Dicembre 2000." In Guido Borio, Francesca Pozzi, and Gigi Roggero, eds. *Futuro anteriore. Dai «Quaderni rossi» ai movimenti globali: ricchezze e limiti dell'operaismo italiano*. Roma: DeriveApprodi.

Armano, Emiliana, Devi Sacchetto, and Steve Wright. 2013. "Coresearch and Counter-Research: Romano Alquati's Itinerary Within and Beyond Italian Radical Political Thought." *Viewpoint Magazine,* Worker's Inquiry (3). Accessed July 29, 2014. http://viewpointmag.com/2013/09/27/coresearch-and-counter-research-romano-alquatis-itinerary-within-and-beyond-italian-radical-political-thought/.

Art is Open Source. 2011. VersuS. Accessed October 15, 2014. http://www.artisopensource.net/projects/versus-the-realtime-lives-of-cities.html.

Baudrillard, Jean. 1995. *Simulacra and Simulation*. Trans. Sheila Faria Glaser. Ann Arbor: University of Michigan Press.

Blas, Zach. 2014. "Facial Weaponization Suite." Accessed September 10, 2014. http://www.zachblas.info/about/.

Blas, Zach. 2013. "Informatic Opacity." *Journal of Aesthetics and Protest* 9. Accessed September 10, 2014. http://www.joaap.org/issue9/zachblas.htm.

Borio, Guido, Francesca Pozzi, and Gigi Roggero. 2002. *Futuro anteriore. Dai «Quaderni rossi» ai movimenti globali: ricchezze e limiti dell'operaismo italiano*. Roma: DeriveApprodi.

Conti, Antonio. 2001. "Inchiesta come metodo politico." *Posse* (2/3): 23–30.

Chun, Wendy Hui Kyong. 2011. *Programmed visions software and memory*. Cambridge, MA: MIT Press.

Combes, Muriel. 2013. *Gilbert Simondon and the Philosophy of the Transindividual*. Cambridge, MA: MIT Press.

Costanza-Chock, Sasha. 2012. "Mic Check! Media Cultures and the Occupy Movement." *Social Movement Studies* 11(3/4): 375–85.

Elmer, Greg and Andy Opel. 2014. *Preempting Dissent Policing the Crisis*. Documentary, Canada, 41 min.

Galloway, Alex. 2012. *The Interface Effect*. Cambridge, UK: Polity.

Gerbaudo, Paolo. 2012. *Tweets and the streets: social media and contemporary activism*. London: PlutoPress.

Gillespie, Tarleton, Pablo Boczkowski, and Kirsten Foot. 2014. *Media technologies: essays on communication, materiality, and society*. Cambridge, MA: MIT Press.

Guattari, Felix. 1995. *Chaosmosis: an ethico-aesthetic paradigm*. Trans. Paul Bains and Julian Pefanis. Sydney: Power Publications.

Guattari, Felix. 2012. "Towards a post-media era." Accessed September 9, 2014. http://www.metamute.org/editorial/lab/towards-post-media-era.

Guattari, Felix. 2000. *The three ecologies*. London; New Brunswick, NJ: Athlone Press.

Juris, Jeffrey S. 2012. “Reflections on #Occupy Everywhere: Social media, public space, and emerging logics of aggregation.” *American Ethnologist* 39(2): 259–79.

Kramer, Ad, Guillory J. E., and Hancock J. T. 2014. “Experimental evidence of massive-scale emotional contagion through social networks.” *Proceedings of the National Academy of Sciences of the United States of America* 111(24): 8788–90.

Latour, Bruno. 1993. *We Have Never Been Modern*. Cambridge, MA: Harvard University Press.

Lovink, Geert and Miriam Rasch. 2013. *Unlike us reader: social media monopolies and their alternatives*. Amsterdam: Institute of Network Cultures.

Magnet, Shoshana. 2011. *When Biometrics Fail: Gender, Race, and the Technology of Identity*. Durham: Duke University Press.

Mejias, Ulises Ali. 2013. *Off the Network: Disrupting the Digital World*. Minneapolis: University of Minnesota Press.

Occupy Data. 2012. “About Occupy Data.” Accessed October 14, 2014. http://occupy-data.org.

Occupy Data. 2013. “#OccupyData NYC.*”* Accessed October 14, 2014. http://occupydatanyc.org/category/projects/.

Pasquinelli, M. 2014. "Italian Operaismo and the Information Machine." *Theory, Culture & Society.* February 2, 2014. doi: 10.1177/0263276413514117

Renzi, Alessandra. Forthcoming. “From the Fax Machine to Social Media: How Information Shapes Social Movements.” *Interface*.

Simondon, Gilbert. 1989a. *Du mode d'existence des objets techniques*. Paris: Aubier.

Simondon, Gilbert. 1989b. *L'Individuation psychique et collective*. Paris: Aubier.

Simondon, Gilbert. 2006. *L' individuazione psichica e collettiva*. Trans. Paolo Virno. Roma: DeriveApprodi.

Stiegler, Bernard 2013. *What Makes Life Worth Living: On Pharmacology*. Trans. Daniel Ross. Cambridge, UK: Polity.

Stiegler, B., R. Daniel, and S. Arnold. 2011. *The Decadence of Industrial Democracies*. Cambridge, UK: Polity Press.

Stiegler, Bernard and Irit Rogoff. 2010. “Transindividuation.” *e-flux journal* 14(March). Accessed September 30, 2014. http://worker01.eflux.com/pdf/article_8888121.pdf.

Terranova, Tiziana. 2000. "Free Labor: Producing Culture for the Digital Economy" *Social Text* 18: 33–58.

Terranova, Tiziana. 2014. "Red Stack Attack! Algorithms, Capital and the Automation of the Common." *Quaderni di San Precario*. Accessed September 9, 2014. http://quaderni.sanprecario.info/2014/02/red-stack-attack-algorithms-capital-and-the-automation-of-the-common-di-tiziana-terranova/.

van Dijck, J. 2014. "Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology." *Surveillance and Society* 12(2): 197–208.

Virno, Paolo. 2008. "Angels and the General Intellect: Individuation in Duns Scotus And Gilbert Simondon." *Parrhesia* 7: 58–67.

Warner, Michael. 2002. "Publics and Counterpublics." *Public Culture* 14(1): 49–90.

Winnicott, Donald W. 1953. "Transitional Objects and Transitional Phenomena—A Study of the First Not-Me Possession.) *International Journal of Psycho-Analysis* 34: 89–97.

Widgington, David. 2013. "Artéfacts d'un Printemps québécois Archive." Accessed October 14, 2014. http://www.printempserable.net.

Article

new media & society
1–19
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1461444816629469
nms.sagepub.com
SAGE

Civic hacking as data activism and advocacy: A history from publicity to open government data

Andrew R Schrock
University of Southern California, USA

Abstract

The civic hacker tends to be described as anachronistic, an ineffective "white hat" compared to more overtly activist cousins. By contrast, I argue that civic hackers' politics emerged from a distinct historical milieu and include potentially powerful modes of political participation. The progressive roots of civic data hacking can be found in early 20th-century notions of "publicity" and the right to information movement. Successive waves of activists saw the Internet as a tool for transparency. The framing of openness shifted in meaning from information to data, weakening of mechanisms for accountability even as it opened up new forms of political participation. Drawing on a year of interviews and participant observation, I suggest civic data hacking can be framed as a form of data activism and advocacy: requesting, digesting, contributing to, modeling, and contesting data. I conclude civic hackers are utopian realists involved in the crafting of algorithmic power and discussing ethics of technology design. They may be misunderstood because open data remediates previous forms of openness. In the process, civic hackers transgress established boundaries of political participation.

Keywords

Activism, hacking, hacktivism, open data, politics, transparency

"Civic hackers" participating in creating and modifying digital infrastructure have garnered increased attention over the last 5 years. They are generally described as a more positive-valenced (Newsom, 2013) cousin of more overtly oppositional activists and

Corresponding author:
Andrew R Schrock, 3321 Monogram Ave., Long Beach, CA 90808, USA.
Email: me@aschrock.com

hacktivists (Taylor, 2005). Earliest definitions lauded civic hackers as “white hats” that create technology to foster stronger social bonds, reflecting a libertarian perspective on mutual aid (Crabtree, 2007). Other definitions capture broader notions of civil society. A 2010 study backed by the Open Society Foundation described civic hackers as “deploying information technology tools to enrich civic life, or to solve particular problems of a civic nature, such as democratic engagement” (Hogge, 2010: 10). Simultaneously, federal and local government entities warmed to the notion of collaborating with tech-literate geeks who could create, interpret, and use data. Anthony Townsend (2013) describes civic hackers as being essential change agents in urban environments. Organizations such as Code for America (CfA) rallied support by positioning civic hacking as a mode of direct participation in improving structures of governance. However, critics objected to the involvement of corporations in civic hacking as well as their dubious political alignment and non-grassroots origins. Critical historian Evgeny Morozov (2013a) suggested that “civic hacker” is an apolitical category imposed by ideologies of “scientism” emanating from Silicon Valley. Tom Slee (2012) similarly described the open data movement as co-opted and neoliberalist. Looking past the respective hype and cynicism, where did the progressive bent of civic hacking come from? What does it have to say about the potentials and pitfalls for political participation through and around data?

Civic hacking can broadly be described as a form of alternative/activist media that “employ or modify the communication artifacts, practices, and social arrangements of new information and communication technologies to challenge or alter dominant, expected, or accepted ways of doing society, culture, and politics” (Lievrouw, 2011: 19). Ample research has considered how changes in technology and access have created “an environment for politics that is increasingly information-rich and communication-intensive” (Bimber, 2001). Earl and Kimport (2011) argue that such digital activism draws attention to modes of protest—“digital repertoires of contention” (p. 180)—more than formalized political movements. A similar middle ground focused on shared histories and practices is suggested by Molly Sauter’s work on distributed denial of service (DDOS) attacks, which traces histories of civil disobedience and a nuanced relationship between evolution of tools to support activist goals (Sauter, 2013b). Similarly, the focus of this article is on how the political practices of civic hackers emerged from a particular legal and historical trajectory. Government entities increasingly encourage and foster the civic hacker as an essential part of this system. They started to follow a unitary rather than adversary model of democracy. Yet, I argued that the movement from informational to data can potentially lead to quite different forms of political action. Tracing how open government data emerged in the United States is thus a necessary first step to explain the beliefs participants place in civic hacking and the fraught institutional tensions they must navigate.

Hacker politics

Definitions of “the political,” formalized politics, and political participation are expansive and endlessly debated. Writing on the intersection between hackers and politics has focused on a more confined set of motivations and goals. Hackers are not simply computer super-users. Rather, over time technology has become integral to hackers’ informational practices (Thomas, 2002), material engagement (Jordan, 2008), and use of tools for collective action (Sauter, 2013b). Two overall framings of hackers' engagement with the political have dominated discussion: "hacktivists" or activists who leverage instrumental uses of online technologies for direct political action such as protest and disruption (Jordan and Taylor, 2004), and geographically distributed communities of practice where principles of openness enable forms of political action (Coleman, 2004). Gabriella Coleman (2012a) argues that pragmatism enables action on issues related to informational freedoms and reflects liberal democratic tenets such as freedom of speech. According to Coleman (2004), explicit involvement in "politics" in a formalized sense is distasteful to free and open-source hackers, as it is viewed as "buggy, mediated, and tainted action clouded by ideology" (p. 513). Civic hacking represents a third mode of participation among a group that often explicitly engages with political causes through designing, critiquing, and manipulating software and data to improve community life and infrastructures of governance. Civic hackers therefore have distinct histories, contours, and conflicts from other genres of hackers, even as they share a certain family resemblance (Wittgenstein, 1953).

The civic hacker's institutionally collaborative nature is the foremost difference from other forms of hackers that are more defined as antagonistic (Söderberg, 2010) or subversive (Thomas, 2002). Paying close attention to practices of civic hackers, then, draws attention to possibilities for designing and modifying digital infrastructures that are often overlooked in prognostications about "big data" (Boyd and Crawford, 2012). For example, Zeynep Tufekci (2014) describes "computational politics" where governments and corporations negatively affect communication through tailored messaging, surveillance, and disrupted public spheres. John Cheney-Lippold describes "algorithmic citizenship" produced when the National Security Agency (NSA) constructed and imposed categories of "citizen" and "foreigner" through statistical processes. In each case data was framed as repressive of notions of civil society or enforcing an impoverished or constrictive notion of citizenship. The perspectives of Tufekci and Cheney-Lippold provide valuable insight into how algorithms and data are powerful shapers of modern life. Yet, they leave little room for a different form of algorithmic citizenship that might emerge where individuals desire to reform technology and data-driven processes. As Couldry and Powell (2014) note, models of algorithmic power (Beer, 2009; Lash, 2007) tend to downplay questions of individual agency. They suggest a need to "highlight not just the risks of creating and sharing data, but the opportunities as well" (p. 5). We should be attentive to moments where meaningful change can occur, even if those changes are fraught with forces of neoliberalism and tinged with technocracy.

The term "hacker" is a floating signifier, articulated and reinterpreted across communities and institutions. Scholars interested in hackers have generally navigated this definitional slipperiness by attempting to unite hackers under a common principle or researched a particular strand of hackers. Critical historian Doug Thomas (2002) unified hacker groups from the 1980s and 1990s through shared culture. He argued they were a postmodern "subculture that resists incorporation by turning incorporation into opportunity" (p. 152). Tim Jordan and Paul Taylor (2004) explored how "hacktivists" shared modes of political action, leveraging technical skills for explicitly political goals. Tim Jordan (2008) would focused on "the hack" as a uniting factor, arguing that hacking is co-constituted by material engagement with technology and social agency (Neff et al., 2012). McKenzie Wark (2004) argued that hackers gained power in modernity because they are able to abstract property to intellectual property, which is used by "vectoralists" (similar to bourgeoisie) in processes of control and commodification. He positioned hackers as potentially collaborative, as both vectoralists and hackers have modes of exchange that reflect forms of power. Hackers create value through information, while vectoralists benefit financially.

The explosion of hacker politics worldwide presents a challenge to unification narratives. It pushed researchers to be attentive to specific historical, cultural, and political groups. Gabriella Coleman followed an anthropological mode of tracing the lifeworld and rich emic perspectives of free and open-source software (F/OSS) enthusiasts as a geographically distributed collective. She laid bare how an ethic of hacking (Levy, 1984), when used as a motivating force for all hackers, can become a convenient shorthand that disguises as much as it reveals (Coleman, 2012b). Coleman suggested scholars consider the diverse range of ways individuals organize, mobilize, and act politically. This move toward how political notions evolve within specific collectives can also be seen in Nathanael Bassett's (2013) writing on activist hackathons and Molly Sauter's (2014) work with online civil disobedience. Their mode of inquiry, and the one I follow here, focuses on a specific collective with a specific and traceable history with attendant tools, practices, and tactics.

This article initially traces a history of informational transparency from "information" to "data." The tools of participation by civic hackers are, as with other geeks and hacker cultures (Coleman, 2012a; Kelty, 2008), rooted in legal frameworks enabling the free flow of information. What we might call early civic hackers came from journalism and law, initially motivated by practices and goals of access to information (Kennedy, 1978). Later, the growing awareness of the Internet sparked interest in digital transparency, or what Lessig (2009) terms "naked transparency." Open government data became instrumentally and ideologically enabled by notions of transparency and specific legal tools such as the Freedom of Information Act (FOIA). Yet, the natural equating of "openness" or government transparency (Hood and Heald, 2006) with accountability increasingly became dubious (Tkacz, 2012). The move to "open data" was often an imperative that didn't make clear where the levers were for social change that benefited citizens (Lessig, 2009). Still, I argue that civic hackers are often uniquely positioned to act on issues of public concern; they are in touch with local communities, with technical skills and, in many cases, institutional and legal literacies. I conclude by connecting the open data movement with a specific set of political tactics—requesting, digesting, contributing, modeling, and contesting data.

Origins of open government data

Doug Thomas began *Hacker Culture* (2002) by outlining the culture of secrecy in the United States as a way to understand both the resistant nature of hackers and why they were vilified. Hacking is still deeply coupled with technology through particular historical trajectories (Coleman, 2012b; Jordan, 2008; Thomas, 2002). Narratives of hackers as deviants were desirable because they solidified support for the government (Nissenbaum, 2004). On one hand, “civic hackers” are welcomed in the current day because the fragility of government makes it increasingly necessary to recognize and invite their labor (Gregg, 2014). On the other, civic hacking is still influenced by a progressive political subjectivity among participants enabled by a specific historical lineage of openness. To provide context for the political participation of civic hackers, I briefly outline a historical arc that enabled the civic hacker as well as critical figures and debates around open government data. Informational freedoms have been given tangibility by laws and practices that enable its flow and utility within particular systems, such as journalism.

This history isn’t intended to be a review of the deployment of technology for civic purposes, which has been covered elsewhere (Goldsmith and Crawford, 2014; Goldstein and Dyson, 2013; Townsend, 2013). Neither is it a philosophical exploration of openness (Birchall, 2012; Tkacz, 2012) or espousing the benefits of transparency in government (Hood and Heald, 2006). Rather, this article is oriented around the movement of specific legal frameworks for open information in the United States and practices of actors in engaging with open government data that enabled what is now termed “civic hacking.” Open government data therefore provides a specific lineage to compare and contrast with the multitude of “open” concepts in circulation, including open standards (Russell, 2014), open source (Coleman, 2012a), and open systems (Kelty, 2008: Chapter 5). While the open government data movement in the United States is kin to these other lineages, it is conceptually and historically distinct. It also bears mention that this history is confined to the United States. Practices with openness have been differently interpreted by grassroots participants across the world, from Asia (Lindtner, 2012) to the global South (Chan, 2013).

Early informational openness: from sunlight to flashlight

Transparency has a variety of definitions, but at its core refers to “the degree to which information is available to outsiders that enables them to have informed voice in decisions and/or to assess the decisions made by insiders” (Florini, 2007: 5). What we now call transparency has its roots in progressive-era notions of “publicity” where business was performed in public. In 1902’s *What is Publicity?* political professor Henry Adams described publicity as “an essential agency for the control of trusts” (p. 895). Woodrow Wilson, long a campaigner for governmental and financial reform, in 1918 beginning his “fourteen points” memo for World War I by calling for an “open convent of peace, openly arrived at” where “diplomacy shall proceed always frankly and in the public view.” Justice Louis Brandeis, with his well-worn aphorism that “sunlight is the best disinfectant” that would inspire the Sunlight Foundation, very much supported and informed Wilson’s position, particularly as discussed in a 1913 Harper’s article and successive book *Other People’s Money.* Brandeis believed that bankers’ compensation should be publicly disclosed to encourage investors to negotiate more reasonable terms. This “full disclosure” would provide information that enables the system to function more efficiently.

By the 1920s, the meaning of publicity had shifted from a universal notion that “sunshine” would bring about smooth functioning and encourage trust toward something more nefarious. Wariness about publicity emerged from the public becoming leery of mass communication. The First World War drew attention to how communication could be wielded to push specific opinions. The publicist became a professional occupation. Progressive Walter Lippmann (1922) famously described the "publicity man" that is "censor and propagandist, responsible only to his employers, and to the whole truth responsible only as it accords with the employers' conception of his own interests" (p. 218). In 1928's *Propaganda* and a series of public debates, Edward Bernays defended the role of the publicist as necessary for the smooth functioning of a democracy. The definition of "publicity," previously lauded, acquired negative connotations of promoting a viewpoint that more benefitted the status quo than provided information that enabled individuals to make rational choices. Journalism historians Stoker and Rawlins (2005) extended Brandeis' metaphor of light to describe this as a move from "searchlight to flashlight"—a narrower and less powerful beam that only illuminated what corporations wanted. Yet they also critiqued progressive beliefs that publicity was synonymous with purification and backers such as Adams for placing "too much faith in information's power to produce public action" (p. 186). The struggle for meaningful social change shifted toward obtaining "correct" publicity, and working within the system also led to an unintended consequence: the progressive movement became a training ground for publicists.

The FOIA

Post-World War II, the United States concentrated power in a secretive national security complex, notably the Central Intelligence Agency (CIA) and NSA. Citizens were left with few methods to obtain information. The most well-known and often utilized legal tool for obtaining information about government operations is the Freedom of Information Act. While this is hardly the only method we might connect with "open information"—there are local, state and national efforts, as well as public interest groups and scientific uses of data harkening back to the 1960s—it is the one that has most closely informed the emerging ecosystem around open government data. For example, radical transparency web advocates such as Carl Malamud in the 1990s used FOIA requests to make large caches of information publicly available for free. The replacement of FOIA requests by open data is still touted by data platforms such as Socrata as a savings of labor and money (Quigg, 2014). Participants in civic data hackathons conceptually connect open data with accountability, to "keep city hall honest" as one participant put it.

In *Advocates of Openness*, George Penn Kennedy (1978) described the freedom of information movement that began after World War I and led up to the creation of the FOIA in 1965. The American Society of Newspaper Editors (ASNE) led a sequence of policy statements in the late 1940s centering on the importance to journalism as a profession to freedom of information, culminating in their soliciting Harold L. Cross to write *The People's Right to Know.* Cross concluded that "there is no enforceable legal right in public or press to inspect any federal non-judicial record." This text was published in 1953 and circulated mostly at the federal level, promoting the idea of freedom of information within and outside of journalist circles as being beneficial to the public good. Congressman John Moss' commission then garnered the attention for President Johnson to sign FOIA into law on 4 July 1966.

Perhaps the most surprising facet of this movement now is its small size. Kennedy (1978) notes that "in the first 10 years … the movement got its impetus from the efforts of a tiny handful of men" (p. 40) in legislature and journalism. Johnson wasn't particularly enthusiastic about signing, and the executive branch pushed for provisions permitting the withholding of information for a wide range of exemptions. FOIA was only strengthened after Nixon's resignation in 1974 when it was amended over Ford's veto. Even afterward, it didn't match with the complete vision of ASNE. At times requests to comply stretched well beyond 10 days and requests are still frequently denied. Despite claiming to embrace openness, Obama's presidency has been notoriously secretive, with record numbers of whistle-blower cases and FOIA requests being denied or censored. Despite FOIA's flaws, implementing a philosophical stance of Americans' "right to information" was notable for several reasons. First, FOIA provided accessible tools to put abstract ideas into practice. Everyday citizens started to attach various political notions to these activities. Second, information flowed into a journalistic ecosystem that was prepared to process and interpret it for everyday citizens. Information obtained through FOIA was being interpreted in stories that changed public opinion (Leff et al., 1986). Third, ability for individuals to request information led to alternate uses for activists, public interest groups, and non-profit organizations.

Late informational transparency

A small cadre of journalists organized and lobbied for legal reforms to cement the concept that citizens had a right to information produced by government entities. FOIA produced a multitude of important stories that changed the course of history, and informational transparency became a mode of political activism (Figure 1). In the early 1990s, government transparency became something of a fad, spawning similar legislation overseas. Leading transparency activists started to view the potential of the Internet for increasing accessibility as a natural extension of this freedom of information movement. The efforts of Carl Malamud and the Sunlight Foundation applied right to information principles to the Internet, facilitating public access to vital information on law and government before individuals requested it. They took a more overtly ecological perspective on openness where information could be integrated into as-yet unforeseen processes. For example, Carl Malamud put Securities and Exchange Commission filings online in 1993. In 2006, Mike Klein founded the Sunlight Foundation, taking its name from Louis Brandeis' well-worn aphorism. They began a well-heeled effort to use "21st-century information technology and Web 2.0 energy" to improve access to information about elected officials.

Lessig's notion that code had regulatory capacities was influential on this early stage of defining open data. Yet, memorable phrases such as "code is law" and quotes such as "to the extent that code becomes open, government's power is reduced" (Lessig, 2006: 152) were often misinterpreted on face value. His stance was not cyberlibertarian (Barbrook and Cameron, 1996). As his successive refutation of transparency in this shift toward open data indicates (Lessig, 2009), he was quite concerned about efforts with software becoming distanced from tangible outcomes. Lessig might regarded as a hacker in the mold of Tim Jordan (2008), taking a progressive perspective on how we might regulate technologies—alongside laws, norms, and markets—that affect behavior.

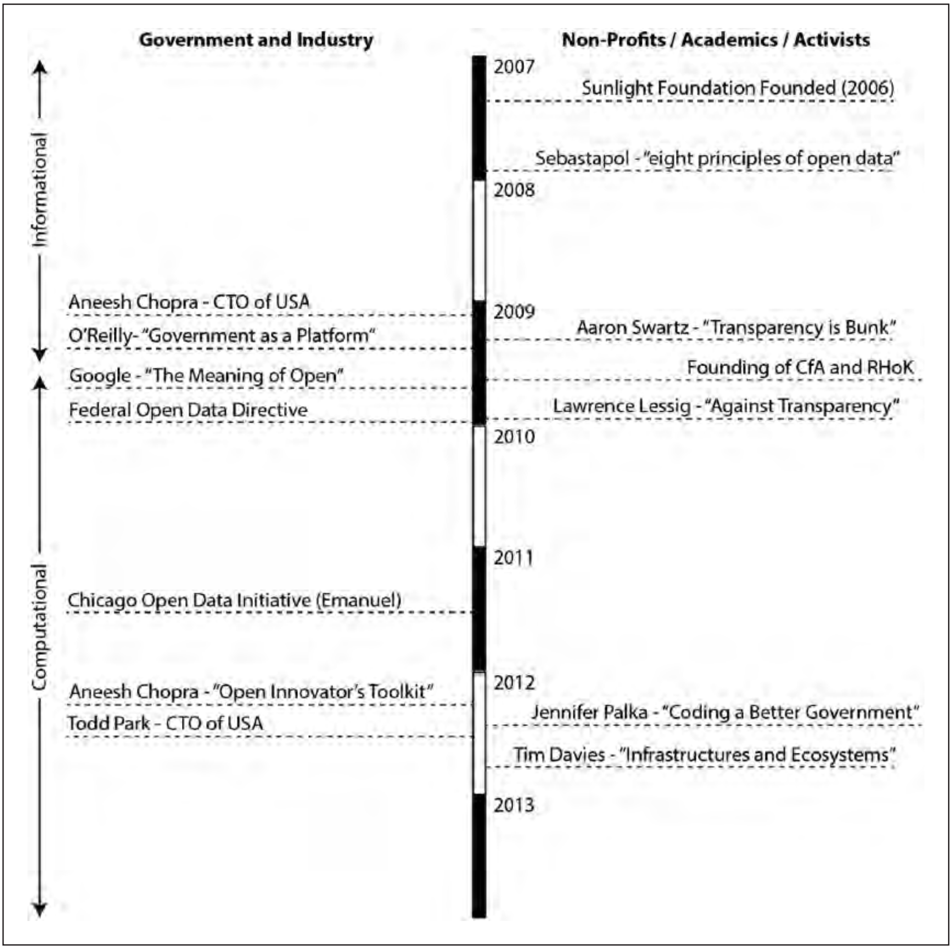

Government and Industry
Non-Profits / Academics / Activists
Informational
Computational
2007
Sunlight Foundation Founded (2006)
Sebastapol - "eight principles of open data"
2008
Aneesh Chopra - CTO of USA
2009
O'Reilly- "Government as a Platform"
Aaron Swartz - "Transparency is Bunk"
Founding of CfA and RHoK
Google - "The Meaning of Open"
Lawrence Lessig - "Against Transparency"
Federal Open Data Directive
2010
2011
Chicago Open Data Initiative (Emanuel)
2012
Aneesh Chopra - "Open Innovator's Toolkit"
Jennifer Palka - "Coding a Better Government"
Todd Park - CTO of USA
Tim Davies - "Infrastructures and Ecosystems"
2013

Figure 1. Timeline of openness from late informational transparency to computational open data.

Definitions of "open data" were successively codified through informal standards that were influential to successive implementations of open data initiatives at the federal and municipal levels.[1] In 2005, the Open Knowledge Foundation produced what they term "the open definition," which could be applied to content, data, and information. Carl Malamud and Tim O'Reilly, backed by the Sunlight foundation, assembled a team of thinkers in 2007 at Sebastopol, north of San Francisco. Participants included Lawrence Lessig and a young Aaron Swartz. The open data definition drafted at Sebastopol describes data's completeness, primacy, timeliness, ease of physical and electronic access, machine readability, non-discrimination, use of commonly owned standards, licensing, permanence, and usage costs. This description made it clear what the properties of data were, even as outcomes, fitting with an open-source model, were more ambiguous. Still, Tim Davies' (2010) taxonomy of open government data of this period includes mechanisms based on political participation, collaboration, and choice. Vocabulary from Sebastopol made its way into the 2013 executive order "Making Open and Machine Readable the New Default for Government Information." On the level of municipal governments in particular, the movement from information to data focused on new uses that emphasized collaboration and utility over accountability (Yu and Robinson, 2012), signaling what I term the computational shift of openness.

The computational shift

The computational shift of open government data refers to the move from governments fulfilling information requests to automatically releasing data to fulfill a range of more speculative uses. While promises about the Internet (Morozov, 2013b) encouraged this move, so too did notions of open government from previous decades. For example, David Osborne's notion of "reinventing government" involved hallmarks familiar to open data initiatives: "catalytic" public–private relationships, connecting with communities, and decentralized collaboration (Osborne and Gaebler, 1992). In 1995's *What Comes Next* republican deputy director of domestic policy for the Bush Administration Jim Pinkerton, frustrated with New Deal era centralization, identified "bugs" with the "operating system" of bureaucracy to reduce bloat. For obvious reasons, civic hackers have gained traction among government officials by promising to streamline processes for allocating funding. Others focused on generating value from new sources. The notion of "network governance" proposed that government officials are responsible for "producing public value rather than managing inputs," assembling packages that are "most useful for the customer" (Goldsmith, 2005: 57). The emphasis on innovation is also visible in Tim O'Reilly's (2010) influential notion of "government as platform," which positioned systems of governance as being similar to technical systems, subject to constant observation and tweaking to improve inefficiencies. He applied a biological model to government, where "information produced by and on behalf of citizens is the lifeblood of the economy and the nation" (O'Reilly, 2010: 14).

Definitional tensions during the transition from informational to computational modes can be seen in writing of both enthusiastic backers of open data and more wary law-based analyses of critics. Jeffrey D. Rubenstein (2013), CEO of procurement analytics platform Smartprocure, claims that release of FOIA information in data form can be "more than transparency; it can be the basis for true collaboration" (p. 81). Crowdsourcing is often used as a metaphor for open data initiatives with emergent and vaguely defined goals of collaboration rather than specific ones (Brabham, 2013). Open data came increasinly referred to an ecosystem of production rather than accountability. In *The New Ambiguity of Open Government*, Harlan Yu and David Robinson (2012) note that open data signals a movement toward "politically neutral public sector disclosures that are easy to reuse, even if they have nothing to do with public accountability" (p. 178). An ecosystemic metaphor was then familiar to e-government practitioners aligned with David Osborne, as well as to urban planners (Light, 2009).

Not all Sebastopol supporters were on board with this shift. Lawrence Lessig and the late Aaron Swartz, important early supporters of the digital transparency movement, were leery about the detaching of informational transparency and accountability. Lessig doubted whether the "naked transparency movement" provided the context for citizen decision-making because it makes unreasonable expectations on their existing knowledge and time. For example, in their analysis of transparency efforts, Fung et al. (2007) found that the most effective policies provided data on which the public can make informed choices. Swartz objected because transparency for transparency's sake shifted labor from government entities to everyday citizens, and the connection of transparency to accountability had been irrevocably altered: "the pipeline of leak to investigation to revelation to report to reform has broken down." In their opinion, flows of information became detached from their uses to gain leverage against corruption.

In 2009, open data received recognition from the federal level, and many state and local efforts followed. In a general sense, optimism among the tech-friendly Obama and his supporters, who used in his first campaign, buoyed interest in open data. Obama appointed Aneesh Chopra as the first chief technology officer (CTO) of the United States and released executive orders defining open government in 2009 as entailing transparency ("conduct its work more openly and publish its information online"), participation (in decision-making), and collaboration (both internal and external). In 2013, he emphasized machine-readable data as a "default for government." While these orders were more a set of progressive clarifications, they were mirrored by similar orders on the municipal level in Chicago (2011), New York (2012), and Los Angeles (2013).

Proponents of open government data tend to view municipalities as being more receptive to these possibilities. In *The Metropolitan Revolution*, Bruce Katz and Jennifer Bradley, beating a decades-old drum, describe the federal government as "partisan, hopelessly fragmented and compartmentalized, frustratingly bureaucratic, and prescriptive." The non-profit Code for American (CfA) was also founded at this time with goals of increasing government effectiveness and transparency and has been effective in connecting volunteers to sympathetic municipal governments and funding. The organization has become something of a pipeline for young talent to quickly rise up the ranks of government; in 2014, 26-year-old Abhi Nemani, ex-co-executive director of CfA, was selected to be the chief data officer for Los Angeles. CfA founder Jennifer Palka became deputy CTO of the United States. Open data, like open information before it, promised fixes for bureaucratic problems and leveling power asymmetries (Fenster, 2012). Municipal governments strapped for funds and in dire need of more efficient frameworks have, of course, welcomed the message that open government data can alleviate time-consuming FOIA requests, make services easier for residents to use, and drive hackathons as a form of public outreach.

The first National Day of Civic Hacking in 2013 cemented the concept in the public mind, even as it revealed hackathons as an odd overlap between do it yourself (DIY) culture and Silicon Valley modes of problem-solving (Gregg and DiSalvo, 2013). Peeking through the cracks was enthusiasm that open data could be a form of financial innovation. This speculation was solidified in a McKinsey Institute report that claimed potential revenues of three trillion dollars in open data across multiple markets (Manyika et al., 2013).

Data activism and advocacy

Civic data hackers emerged from this historical trajectory to operate through a range of data-driven political modes, or what I term "data activism and advocacy," to bring about systematic change. Their collaborative nature is particularly important to prompt academics to revisit debates about how critical and empirical traditions have historically spent insufficient time evaluating possibilities for institutional reform (Melody and Mansell, 1983). Data activism and advocacy ranges from civic engagement (Putnam, 2001) to more oppositional activism (Jordan, 2001). In this sense, it is a specific association of technologically mediated participation with particular political goals (Lievrouw, 2011) resulting in a wide range of tactics. Although open government data is still evolving and is constrained by predictions for economic growth and self-regulation, I argue it enables civic hackers to participate in civic data politics. This is particularly important because data-driven environment is often distanced from providing individuals a sense of agency to change their conditions (Couldry and Powell, 2014). Data activism and advocacy can take place through organizing on related topics, online through mediated data repositories such as Github, and in-person events such as hackathons. The evolution of open government data has left traces on the beliefs individuals place in data as a particular object or system. *Digesting* and *requesting* are modes in many ways carried over from the FOIA and tend to be aligned with anticipated uses of data by government. *Contributing, modeling*, and *contesting* stem from residents leveraging possibilities of open data and software production to attempt to alter process of governance.

Requesting

Open data advocates initially focused on the imperative of, as one civic hacker put it at an open data summit, "sucking [data] out of its database and exposing it." Civic hackers widely view machine-readable data as more useful because it drives a wider variety of potential uses, even as the shift from informational uses raises the bar to the literacies required to interpret it. In civic hackathons, knowledge of government operations was as useful as technical knowledge. Ironically, many data sets considered "open" are not easily findable by outsiders, as government employees often fail to consider how the sets might be found or their value to others. Requesting, as an extension of informational transparency, is still necessary. The FOIA enabled individuals to obtain information from within government and feed into a receptive system of journalists and organizations. Connecting informational and data paradigms can be made on an infrastructural level as well, as platforms like Socrata tout their ability to provide a cost savings through predictively fulfilling FOIA requests and preventing duplication. The concept of "open data" is built on legal frameworks and numerous open data initiatives at the federal, state, and local levels. Processes that at one point were reliant on FOIA have become increasingly automated, drawing suspicion of one-enthusiastic proponents who connected transparency with accountability. Carl Malamud's desire to make government information radically accessible through the web is more than a symbolic desire. After all, "government transparency cannot be defined as only the information that governments deign to share with the public" (Sifry, 2011: 186). Requesting data still holds a place in the repertoire of civic hackers.

Digesting

Hackers have long been viewed as experts capable of applying technical knowledge to bring about systemic change (Söderberg, 2010). Digesting is a process of interpretation and use that was previously served in an informational fashion by journalists in pursuit of a story. The move toward open data raises the challenge of apprehending the meaning and possible uses of data and hence the need for civic hackers (Townsend, 2013). Civic hackers can be viewed as kin to the right to information movements (Beyer, 2014) and operations such as Wikileaks (Sifry, 2011), although they embrace a more overtly local and ecological model (Light, 2009). Michael Schudson (1999) argues that the progressive-era model of the "informed citizen" dissuaded participation when compared to the previous rowdy spectacle of the party era and was weakened during the post-war period. He proposes that "monitorial citizens" act as a watchdog for specific issues, ready to take action. From this perspective, civic hackers could be considered a monitorial elite, watching data streams and processes of algorithmic regulation for injustices and engaging directly with local politics. "The local" operates as a point of collaboration (Dunbar-Hester, 2013) and point of entry for geeks to engage with neighborhood issues. In retrospect, the threat of the online community debates of the 1990s (Baym, 1995) was not that online communities could be as meaningful as offline communities. Rather, as Craig Calhoun (1998) notes, disputes over community as a particular category threatens to distract from a general focus on solidarity by activating "social bases of discursive publics that engage people across lines of basic difference in collective identities" (p. 374). A mutable, popularized hacker identity may have this potential, capable of processing and interpreting abstract systems of regulation.

Contributing

Stephen Goldsmith and Susan Crawford (2014) argue that resident voice should be heard through data and around issues concerning data. Contributing to data sets is often paired with local grassroots use of mobile devices and amplification of local knowledge (Gordon et al., 2011). Using an app that contributes to a shared data resource provides a low barrier to participation. The most popular apps to date have been highly instrumental ways to request services to fix city infrastructure, such as SeeClickFix, a platform that lets residents take pictures of issues that need repair, that are delivered to the appropriate city department as an actionable item. We might think of this as a base-level civic act similar to picking litter off the ground or painting over graffiti. Other activities are thicker modes of participation by generating data or metadata. The primary effort of the 2014 *CodeAcross* effort was to map existing sources of open data. The leader of the event, D.W. Ferrell, described "our role as citizens is to complement" efforts by the government and organizations such as CfA. Contributing to data repositories served purposes for multiple stakeholders: the group created a resource, communicated how well (or poorly) municipalities were releasing data, and came together around ways to digest data. Modes of data activism and advocacy are often interwoven, and it might be rare for groups or individuals to pursue just a single one.

Modeling

Modeling refers to using code and open data to create working or partly working prototypes. Civic hackers such as Jacob Solomon (2014) view apps and software as examples of effective process, driving the idea that government services might be more just, humanistic, and effective in reaching residents. That "hackers" can model beneficial process disrupts the often presumed subversive nature of hacking as much as it does easy assumptions about a Foucaultian notion of governmentality. Prototypes act as working evidence to lobby for changing government process, particularly those that improve digital infrastructure or direct communication with citizens. The capability of code to act as a persuasive argument has long been noted, and modeling can produce charged debates about the very meaning of "civic." For example, The Detroit Water Project connects individuals unable to pay their water bill with those willing to pay it. The website sparked a conversation that spread through news and social media: how might we take collective action to help those deprived of a basic quality of life? How is the deteriorating infrastructure of Detroit a result of larger geographic and economic conditions? On a level of hackathons, prototypes can be speculative (Lodato and DiSalvo, in press) rather than an "outcome," revealing conflicting notions of "civic tech" (Shaw, 2014).

Contesting

Contesting refers to the creation of crowdsourced data or prototypes for not yet existent uses for data. It is similar to modeling but with an oppositional rather than persuasive tone. For example, the shooting of Michael Brown in Ferguson, Missouri, led to protests paired with data activism. Jim Fischer, who has long noted that there are limited national-level data on officer-involved shootings, suggested that data sets on officer-involved shootings be crowdsourced. In other words, data could be aggregated by individuals who would, where necessary, request additional documentation. The various levels of government are "transparent" about these incidents to a point; the data provided only neutral and in some cases outright false descriptions. Deploying a data-driven vision—what is missing? What can't we see in the data and why?—was a rallying cry for participation to bring about increased accountability. In response to Ferguson, a group of African-American youth created an app "Five-O" to rate local police, receiving national-level attention. It was a simple app created in the spirit of online apps rating goods and services such as Yelp. The juxtaposition of rating and lack of trust highlighted an alternate definition of "safety" that was markedly absent in the previous example of using government data on crime to make residents safe. One relied on a government-sponsored vision of "safety" while the other sought to foster increased accountability among law enforcement officers.

Conclusion: civic hackers as utopian realists

Data activism and advocacy provides a mode of participation in digital infrastructures that debates and confronts the politics of technology for governance. Proponents of open data revive progressive-era claims of transparency as "sunlight" where open data leads to accountability. I suggest that this canard shouldn't lead us to entirely dismiss the movement as encapsulating meaningful modes of political engagement. "Hacker" poorly describes the emergence of civic hackers' tactics from informational transparency, particularly as they are not "white hats" in the historical sense of reformed hackers joining the workforce (Sauter, 2013a), nor simply an inversion of "black hat" information security experts. Rather, civic hackers seek to ease societal suffering by bringing the hidden workings of abstract systems to light and improve their functioning. Part of the academic discomfort with recognizing civic hacking might stem from their activities cutting across political categories that have traditionally been passionately defended: unitary and adversary, citizen and consumer, horizontalist and institutionalized, and prefigurative and strategic.

Civic hackers might be most appropriately described as utopian realists (Giddens, 1990: 154), a term Giddens employed to capture how assuaging negative consequences in a risk society required retaining Marx's concern of connecting social change to institutional possibilities while leaving behind his formulation of history as determining and reliance on the proletariat as change agents. He positioned utopian realists as sensitive to social change, capable of creating positive models of society, and connecting with life politics. Giddens received criticism for applying the term to movement-based politics. Might "utopian realist" be applicable to the practices of civic hackers, intertwined with particular repertoires, technologies, and affective publics? McKenzie Wark (2014) suggests that the relationship between utopian and realist might be mutually constitutive rather than dialectical. He re-frames utopia as a realizable fragment or diagram that re-imagines relations. From this perspective, civic hacking gets traction not because they were ever intended to be the sole "solution" to a problem, but they are ways of acting and creating that are immediately apprehensible. Prototypes capture the imagination because they are shards of a possible future and can be created, modified, and argued about (Coleman, 2009).

The rise of civic hackers from informational transparency and the inversion of the negative valence of "hacker" were not by chance. As the history outlined here suggests, their politics are the result of the evolution of informational openness and move toward "big data," leading to new possibilities for collaboration even as it slides toward neoliberalism. The perspective that there is civic potential in freeing data has been enabled by a particular historical moment with a rise of alternative modes of political engagement outside normative roles (Schudson, 1999), increased numbers of technically literate with free time (Neff, 2012), and a desire among governments to re-frame "civic hackers" using a positive valence (Gregg, 2014). To civic hackers, changing one part of a system drives humanistic design processes and services for those in need, influencing other parts. While there is not sufficient time to discuss this more systematic perspective, the appeal of open government data is that it acts as a soft form of power to bring out the positive qualities of cities. Despite their fetish for speed (in rapid production and hackathons), civic hackers often act as the "slow food movement" of digital political action, embracing local sourcing, ethical consumption, and pleasure of community work. Civic hackers are hardly responding to a new narrative of technological change. Matthew Wisnioski (2012) notes that the normative framing of rapid technological change leading to institutions being unable to catch up has been surprisingly stable over the last 60 years. Civic hackers thus speak against those who propose that the application of technology to politics produces a meta-category of activist.

Civic hackers tackle a difficult and timeworn problem: participation in humanistic technology design. Perhaps their greatest achievement is showing how "publics suited to renewed discussion about technological choices and policies might be constituted" (Winner, 1992: 355). Currently, open government data and its emergent ecosystem jostle uncomfortably between liberal democratic and cyberlibertarian perspectives; open data can be viewed as an opportunity to help communities be more justly governed or a justification for complete government disintermediation. A growing awareness of these divergent views has led to civic hackers actively debate ethics of representation and how to articulate shared political values (Shaw, 2014). As should be clear, I don't view civic hackers as simply pragmatists who have adopted a cybernetic ideology embedded in the Internet (Morozov, 2014). To put Giddens in conversation with Morozov, the threat of civic hackers is not that they naively employ "solutionism." Quite to the contrary, they debate ethics of technology design, seek collaborations with local organizations, and attempt to re-think how government services might be more sensitive to resident needs. The more pressing threat is that a fear of solutionism and neoliberal connotations of "open data" together might dissuade political participation. Systemic social disparities are often intractable. The route to alleviate them has never been detachment or abandonment. Looking forward, we should pay attention to how data activism and advocacy might result in meaningful systematic change beyond the usual claims of "transparency." To fulfill the possibilities for meaningful social change hinted at in their history, civic hackers might have to coordinate around specific mechanisms for change and articulate a deeper sense of democracy than the language of technology provides.

Acknowledgements

The author would like to thank the two anonymous reviewers and Morgan Currie, Nathalie Marechal, and Brett Shears, who provided feedback on earlier drafts. This article was further assisted by feedback at events where these ideas were worked out, namely, the civic hacking panel at the 2014 Digital Labor conference with Carl DiSalvo, Melissa Gregg, Max Liboiron, Lilly Irani, and Thomas Lodato; and a panel on hacking at the 2014 International Communication Association conference with Molly Sauter, Sebastian Kubitschko, and Nick Couldry. Finally, the author would like to express appreciation for Henry Jenkins and the Civic Paths research group, which provided invaluable early feedback on a presentation and series of blog posts. Without a supportive home and appreciative ears, this article simply would not exist.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

[1] Although I am focused on the United States, it bears mention that the United Kingdom preceded the United States in several initiatives and uses of open data—see Tim Davies' fine work on this subject: http://www.opendataimpacts.net/report/

References

Barbrook R and Cameron A (1996) The Californian ideology. *Science as Culture* 6(1): 44–72.

Bassett N (2013) *The conscientious hacker: an ethnography of identity and community among hackathons*. Master of Arts Thesis, New School, New York.

Baym NK (1995) The emergence of community in computer-mediated communication. In: Jones S (ed.) *Cybersociety: Computer-Mediated Communication and Community*. Thousand Oaks, CA: SAGE, pp. 138–163.

Beer D (2009) Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society* 11(6): 985–1002.

Beyer JL (2014) The emergence of a freedom of information movement: anonymous, Wikileaks, the pirate party, and Iceland. *Journal of Computer-Mediated Communication* 19(2): 141–154.

Bimber B (2001) Information and political engagement in America: the search for effects of information technology at the individual level. *Political Research Quarterly* 54(1): 53.

Birchall C (2012) Introduction to "secrecy and transparency": the politics of opacity and openness. *Theory, Culture & Society* 28(7–8): 7–25.

Boyd D and Crawford K (2012) Critical questions for big data. *Information, Communication & Society* 15(5): 662–679.

Brabham DC (2013) *Crowdsourcing*. Cambridge, MA: MIT Press.

Calhoun C (1998) Community without propinquity revisited: communications technology and the transformation of the urban public sphere. *Sociological Inquiry* 68: 373–397.

Chan AS (2013) *Networking Peripheries: Technological Futures and the Myth of Digital Universalism*. Cambridge, MA: MIT Press.

Coleman G (2004) The political agnosticism of free and open source software and the inadvertent politics of contrast. *Anthropological Quarterly* 77(3): 507–519.

Coleman G (2009) Code is speech: legal tinkering, expertise, and protest among free and open source software developers. *Cultural Anthropology* 24(3): 420–454. Available at: http://doi.wiley.com/10.1111/j.1548-1360.2009.01036.x

Coleman G (2012a) *Coding Freedom: The Ethics and Aesthetics of Hacking*. Princeton, NJ: Princeton University Press.

Coleman G (2012b) Phreaks, hackers and trolls. In: Mandiberg M (ed.) *The Social Media Reader*. New York: New York University Press, pp. 99–119.

Couldry N and Powell A (2014) Big data from the bottom up. *Big Data & Society* 1(2): 1–5.

Crabtree J (2007) Civic hacking: a new agenda for e-democracy. *OpenDemocracy*. Available at: http://www.opendemocracy.net/debates/article-8-85-1025.jsp

Davies T (2010) Open Data, Democracy and Public Sector Reform. Available at: http://www.opendataimpacts.net/report/

Dunbar-Hester C (2013) What's local? Localism as a discursive boundary object in low-power radio policymaking. *Communication, Culture & Critique* 6(4): 502–524.

Earl J and Kimport K (2011) *Digitally Enabled Social Change: Activism in the Internet Age*. Cambridge, MA: MIT Press.

Fenster M (2012) *The Transparency Fix: Advocating Legal Rights and Their Alternatives in the Pursuit of a Visible State*. Available at: http://scholarship.law.ufl.edu/cgi/viewcontent.cgi?article=1385&context=facultypub

Florini A (2007) *The Right to Know: Transparency for an Open World*. New York: Columbia University Press.

Fung A, Graham M and Weil D (2007) *Full Disclosure*. Cambridge: Cambridge University Press.

Giddens A (1990) *The Consequences of Modernity*. Stanford, CA: Stanford University Press.

Goldsmith S (2005) Information technology as a tool for government reform. In: Blackstone EA, Bognanno ML and Hakim S (eds) *Innovations in E-Government*. Lanham, MD: Rowman & Littlefield Publishers, Inc., pp. 56–60.

Goldsmith S and Crawford S (2014) *The Responsive City: Engaging Communities Through Data-Smart Governance*. San Francisco, CA: Jossey-Bass.

Goldstein B and Dyson L (2013) *Beyond Transparency: Open Data and the Future of Government*. San Francisco, CA: Code for America press.

Gordon E, De Souza E and Silva A (2011) *Net.Locality: Why Location Matters in a Networked World*. Malden, MA: Wiley-Blackwell.

Gregg M (2014) *Hack for Good: Speculative Labor, App Development and the Burden of Austerity*. Available at: http://twentyfive.fibreculturejournal.org/fcj-186-hack-for-good-speculative-labour-app-development-and-the-burden-of-austerity/

Gregg M and DiSalvo C (2013) The trouble with white hats. *The New Inquiry*. Available at: http://thenewinquiry.com/essays/the-trouble-with-white-hats/

Hogge B (2010) Open Data Study: New Technologies. Available at: https://www.opensociety-foundations.org/sites/default/files/open-data-study-20110519.pdf

Hood C and Heald D (2006) *Transparency: The Key to Better Governance?* Oxford: Oxford University Press.

Jordan T (2001) *Activism! Direct Action, Hacktivism and the Future of Society*. London: Reaktion Books Ltd.

Jordan T (2008) *Hacking: Digital Media and Technological Determinism*. Cambridge; Malden, MA: Polity Press.

Jordan T and Taylor PA (2004) *Hacktivism and Cyberwars*. London: Routledge.

Kelty CM (2008) *Two Bits: The Cultural Significance of Free Software*. Durham, NC: Duke University Press.

Kennedy GP (1978) *Advocates of openness: The freedom of information movement*. PhD Thesis, University of Missouri, Columbia, MI.

Lash S (2007) Power after hegemony: cultural studies in mutation? *Theory, Culture & Society* 24(3): 55–78.

Leff DR, Protess DL and Brooks SC (1986) Crusading journalism: changing public attitudes and policy-making agendas. *Public Opinion Quarterly* 50(3): 300–315.

Lessig L (2006) *Code: And Other Laws of Cyberspace*, version 2.0. New York: Basic Books.

Lessig L (2009) Against transparency. *The New Republic*. Available at: https://newrepublic.com/article/70097/against-transparency

Levy S (1984) *Hackers: Heroes of the Computer Revolution*. Garden City, NY: Anchor Press/Doubleday.

Lievrouw L (2011) *Alternative and Activist New Media*. Cambridge: Polity Press.

Light J (2009) *Nature of Cities: Ecological Visions and the American Urban Professions, 1920–1960*. Baltimore, MD: Johns Hopkins University Press.

Lindtner S (2012) Remaking creativity & innovation: China's nascent DIY maker & hackerspace community. Available at: http://www.ics.uci.edu/~lindtner/documents/Lindtner_Remakinginnovation.pdf

Lippmann W (1922) *Public Opinion*. New York: Harcourt, Brace and Company.

Lodato T and DiSalvo C (in press) Issue-Oriented Hackathons as Material Participation. *New Media & Society*.

Manyika J, Chui M, Farrell D, et al. (2013) Open data: unlocking innovation and performance with liquid information. Report, McKinsey Global Institute. Available at: http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information

Melody WH and Mansell RE (1983) The debate over critical vs. Administrative research: circularity or challenge. *Journal of Communication* 33(3): 103–116.

Morozov E (2013a) The meme hustler. Available at: http://thebaffler.com/past/the_meme_hustler

Morozov E (2013b) *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: Public Affairs.

Morozov E (2014) The planning machine: Project Cybersyn and the origins of the big data nation. *The New Yorker*. Available at: http://systemagazin.com/planning-machine-project-cybersyn-origins-big-data-nation/

Neff G (2012) *Venture Labor*. Cambridge, MA: MIT Press.

Neff G, Jordan T, Mcveigh-Schultz J, et al. (2012) Affordances, technical agency, and the politics of technologies of cultural production. *Journal of Broadcasting & Electronic Media* 56(2): 299–313.

Newsom G (2013) *Citizenville: How to Take the Town Square Digital and Reinvent Government*. London: Penguin Press.

Nissenbaum H (2004) Hackers and the contested ontology of cyberspace. *New Media & Society* 6(2): 195–217.

O'Reilly T (2010) Government as a platform. *Innovations* 6(1): 13–40.

Osborne D and Gaebler T (1992) *Reinventing Government: How the Entrepreneurial Spirit Is Transforming the Public Sector*. New York: Addison-Wesley Publishing.

Putnam RD (2001) *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon & Schuster.

Quigg B (2014) 5 government processes replaced by open data. *Socrata*. Available at: http://www.socrata.com/blog/5-government-processes-replaced-by-open-data/ (accessed 6 March).

Rubenstein JD (2013) Hacking FOIA: using FOIA requests to drive government innovation. In: Goldstein B and Dyson L (eds) *Beyond Transparency: Open Data and the Future of Government*. San Francisco, CA: Code for America Press, pp. 81–92.

Russell AL (2014) *Open Standards and the Digital Age*. Cambridge: Cambridge University Press.

Sauter M (2013a) *Kevin Mitnick, the New York Times, and the Media's Conception of the Hacker*. IAMCR. Available at: http://www.iamcr2013dublin.org/content/kevin-mitnick-new-york-times-and-medias-conception-hacker

Sauter M (2013b) "LOIC will tear us apart": the impact of tool design and media portrayals in the success of activist DDOS attacks. *American Behavioral Scientist* 57(7): 983–1007.

Sauter M (2014) *The Coming Swarm: DDOS Actions, Hacktivism, and Civil Disobedience on the Internet*. London: Bloomsbury Academic.

Schudson M (1999) *The Good Citizen: A History of American Civic Life*. Cambridge, MA: Harvard University Press.

Shaw E (2014) Civic wants, civic needs, civic tech. *Sunlight Foundation*. Available at: http://sunlightfoundation.com/blog/2014/09/29/civic-wants-civic-needs-civic-tech/ (accessed 29 September).

Sifry ML (2011) *Wikileaks and the Age of Transparency*. Berkeley, CA: Counterpoint.

Slee T (2012) Why the "open data movement" is a joke. *Whimsley*. Available at: http://whimsley.typepad.com/whimsley/2012/05/why-the-open-data-movement-is-a-joke.html (accessed 1 May).

Söderberg J (2010) Misuser inventions and the invention of the misuser: hackers, crackers and filesharers. *Science as Culture* 19(2): 151–179.

Solomon J (2014) People, not data. *Blogging for America*. Available at: http://www.codeforamerica.org/blog/2014/01/06/people-not-data/ (accessed 6 January).

Stoker K and Rawlins BL (2005) The "light" of publicity in the progressive era: from searchlight to flashlight. *Journalism History* 30(4): 177–188.

Taylor PA (2005) From hackers to hacktivists: speed bumps on the global superhighway? *New Media & Society* 7(5): 625–646.

Thomas D (2002) *Hacker Culture*. Minneapolis, MN: University of Minnesota Press.

Tkacz N (2012) From open source to open government: a critique of open politics. *Ephemera: Theory & politics in organization* 12(4): 386–405.

Townsend AM (2013) *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. London: W.W. Norton & Company Ltd.

Tufekci Z (2014) Engineering the public: big data, surveillance and computational politics. *First Monday* 2014(19): 7.

Wark M (2004) *A Hacker Manifesto*. Cambridge, MA: Harvard University Press.

Wark M (2014) Utopian realism. *Public Seminar*. Available at: http://www.publicseminar.org/2014/10/utopian-realism/ (accessed 28 October).

Winner L (1992) Citizen virtues in a technological order. *Inquiry: An Interdisciplinary Journal of Philosophy* 35(3–4): 341–361.

Wisnioski MH (2012) *Engineers for Change: Competing Visions of Technology in 1960s America*. Cambridge, MA: MIT Press.

Wittgenstein L (1953) *Philosophical Investigations*. New York: Macmillan.

Yu H and Robinson DG (2012) The new ambiguity of "open government." *UCLA Law Review* 59: 178.

Author biography

Andrew R Schrock received a PhD from the Annenberg School for Communication and Journalism at the University of Southern California. He is currently active in the public sector in Southern California as a consultant and volunteer. His research and practice currently focuses on how grassroots groups and governments can effectively and ethically use data to improve civic life. Andrew's research has appeared in the *International Journal of Communication, Information, Communication & Society*, and *New Media & Society*. For more information, please visit his website at aschrock.com.

The Ethnography of Infrastructure

SUSAN LEIGH STAR
University of California, San Diego

This article asks methodological questions about studying infrastructure with some of the tools and perspectives of ethnography. Infrastructure is both relational and ecological—it means different things to different groups and it is part of the balance of action, tools, and the built environment, inseparable from them. It also is frequently mundane to the point of boredom, involving things such as plugs, standards, and bureaucratic forms. Some of the difficulties of studying infrastructure are how to scale up from traditional ethnographic sites, how to manage large quantities of data such as those produced by transaction logs, and how to understand the interplay of online and offline behavior. Some of the tricks of the trade involved in meeting these challenges include studying the design of infrastructure, understanding the paradoxes of infrastructure as both transparent and opaque, including invisible work in the ecological analysis, and pinpointing the epistemological status of indictors.

> Resources appear, too, as shared visions of the possible and acceptable dreams of the innovative, as techniques, knowledge, know-how, and the institutions for learning these things. Infrastructure in these terms is a dense interwoven fabric that is, at the same time, dynamic, thoroughly ecological, even fragile.
>
> —Bucciarelli, 1994, p. 131

> Tell that its sculptor well those passions read
> Which yet survive, stamped on these lifeless things.
> —Percy Bysshe Shelley, 1817

GENERAL METHODOLOGICAL PROBLEMS

This article is in a way a call to study boring things. Many aspects of infrastructure are singularly unexciting. They appear as lists of numbers and technical specifications, or as hidden mechanisms subtending those processes more familiar to social scientists. It takes some digging to unearth the dramas inherent in system design creating, to restore narrative to what appears to be dead lists.

Author's Note: *The author thanks Howie Becker, Geof Bowker, Jay Lemke, Nina Wakeford, and Barry Wellman for helpful comments. This article is for the other members of the Society of People Interested in Boring Things, especially cofounder Charlotte Linde.*

AMERICAN BEHAVIORAL SCIENTIST, Vol. 43 No. 3, November/December 1999 377-391

Bowker and Star (in press) note of the International Classification of Diseases (ICD), a global information-collecting system administered by the World Health Organization,

> Reading the ICD is a lot like reading the telephone book. In fact, it is worse. The telephone book, especially the yellow pages, contains a more obvious degree of narrative structure. It tells how local businesses see themselves, how many restaurants of a given ethnicity there are in the locale, whether or not hot tubs or plastic surgeons are to be found there. (Yet most people don't curl up with a good telephone book of a Saturday night.)

They note that aside from this direct information, indirect readings of such dry documents can also be instructive. In the case of phone books, for instance, a slender volume indicates a rural area; those that list only husband's names for married couples indicate a heterosexually-biased, sexist society.

Historical changes are important in reading these documents. Names and locations of services may change with political currents and social movements. Bowker and Star (in press) note that,

> In the Santa Cruz, California, phone book, Alcoholics Anonymous and Narcotics Anonymous are listed in emergency services; years ago they would have been listed under "rehabilitation" if at all. The changed status reflects the widespread recognition of the organizations' reliability in crisis situations, as well as acceptance of their theory of addiction as a medical condition. Under the community events section in the beginning, next to the Garlic Festival and the celebration of the anniversary of the city's founding, the Gay and Lesbian Pride Parade is listed as an annual event. Behind this simple telephone book listing lies decades of activism and conflict—for gays and lesbians, becoming part of the civic infrastructure in this way betokens a kind of public acceptance almost unthinkable 30 years ago . . . excursions into this aspect of information infrastructure can be stiflingly boring. Many classifications appear as nothing more than lists of numbers with labels attached, buried in software menus, users' manuals, or other references.

Much of the ethnographic study of information systems implicitly involves the study of infrastructure. Struggles with infrastructure are built into the very fabric of technical work (Neumann & Star, 1996). However, it is easy to stay within the traditional purview of field studies: talk, community, identity, and group processes, as now mediated by information technology. There have been several good studies of multiuser dungeons (MUDs), or virtual role-playing spaces, distance-mediated identity, cyberspace communities, and status hierarchies. There are much fewer on the effect of standardization or formal classification on group formation, the design of networks and their import for various communities, or on the fierce policy debates about domain names, exchange protocols, or languages.

Perhaps this is not surprising. The latter topics tend to be squirreled away in semi-private settings or buried in inaccessible electronic code. Theirs is not the usual sort of anthropological strangeness. Rather, it is an embedded strangeness, a second-order one, that of the forgotten, the background, the frozen in place.

Studies of gender bending in MUDs, of anonymity in decision making, and new electronic affiliations *are* important; they stretch our understanding of identity, status, and community. The challenges they present are nontrivial methodologically. How does one study action at a distance? How does one even observe the interaction of keyboard, embodied groups, and language? What are the ethics of studying people whose identity you may never know? When is an infrastructure finished, and how would we know that? How do we understand the ecology of work as affected by standardization and classification? What is universal or local about standardized interfaces? Perhaps most important of all, what values and ethical principles do we inscribe in the inner depths of the built information environment (Goguen, 1997; Hanseth & Monteiro, 1996; Hanseth, Monteiro, & Hatling, 1996)? We need new methods to understand this imbrication of infrastructure and human organization.

As well as the important studies of body snatching, identity tourism, and transglobal knowledge networks, let us also attend ethnographically to the plugs, settings, sizes, and other profoundly mundane aspects of cyberspace, in some of the same ways we might parse a telephone book. My teacher Anselm Strauss had a favorite aphorism, "study the unstudied." This led him and his students to research in understudied areas: chronic illness (Strauss, 1979), low-status workers such as janitors, death and dying, and the materials used in life sciences including experimental animals and taxidermy (Clarke & Fujimura, 1992). The aphorism was not a methodological perversion. Rather, it opened a more ecological understanding of workplaces, materiality, and interaction, and underpinned a social justice agenda by valorizing previously neglected people and things.

The ecological effect of studying boring things (infrastructure, in this case) is in some ways similar. The ecology of the distributed high-tech workplace, home, or school is profoundly impacted by the relatively unstudied infrastructure that permeates all its functions. Study a city and neglect its sewers and power supplies (as many have), and you miss essential aspects of distributional justice and planning power (Latour & Hermant, 1998). Study an information system and neglect its standards, wires, and settings, and you miss equally essential aspects of aesthetics, justice, and change. Perhaps if we stopped thinking of computers as information highways and began to think of them more modestly as symbolic sewers, this realm would open up a bit.

DEFINING INFRASTRUCTURE

> What can be studied is always a relationship or an infinite regress of relationships. Never a "thing." (Bateson, 1978, p. 249)

People commonly envision infrastructure as a system of substrates—railroad lines, pipes and plumbing, electrical power plants, and wires. It is by definition invisible, part of the background for other kinds of work. It is ready-to-hand. This image holds up well enough for many purposes—turn on the faucet for a drink of water and you use a vast infrastructure of plumbing and water regulation without usually thinking much about it.

The image becomes more complicated when one begins to investigate large-scale technical systems in the making, or to examine the situations of those who are *not* served by a particular infrastructure. For a railroad engineer, the rails are not infrastructure but topic. For the person in a wheelchair, the stairs and door-jamb in front of a building are not seamless subtenders of use, but barriers (Star, 1991). One person's infrastructure is another's topic, or difficulty. As Star and Ruhleder (1996) put it, infrastructure is a fundamentally relational concept, becoming real infrastructure in relation to organized practices (see also Jewett & Kling, 1991). So, within a given cultural context, the cook considers the water system as working infrastructure integral to making dinner. For the city planner or the plumber, it is a variable in a complex planning process or a target for repair: "Analytically, infrastructure appears only as a relational property, not as a thing stripped of use" (Star & Ruhleder, 1996, p. 113).

In my own research, this became clear when I did fieldwork over 3 years with a community of biologists, in partnership with a computer scientist who was building an electronic shared laboratory and publishing space for them (Schatz, 1991). I was studying their work practices and traveling to many laboratories to observe computer use and communication patterns. Although we were following the principles of participatory design—using ethnography to understand the details of work practice, extensive prototyping, and user feedback; testing the system in laboratories and at conferences—few biologists ended up using the system. It seemed the difficulty was not in the interface or the representation of the work processes embedded in the system, but rather in infrastructure—incompatible platforms, recalcitrant local computing centers, and bottlenecked resources. We were forced to develop a more relational definition of infrastructure, and at the same time, challenge received views of good use of ethnography in systems development.

We began to see infrastructure as part of human organization, and as problematic as any other. We performed what Bowker (1994) has called an "infrastructural inversion"—foregrounding the truly backstage elements of work practice. Recent work in the history of science (Bowker, 1994; Edwards, 1996; Hughes, 1983, 1989; Summerton, 1994; Yates, 1989) has begun to describe the history of large-scale systems in precisely this way. Whether in science or in the arts, we see and name things differently under different infrastructural regimes. Technological developments move from either independent or dependent variables, to processes and relations braided in with thought and work. In the Worm Community Study, Ruhleder and I came to define *infrastructure* as having the following properties, with examples following each dimension.

Embeddedness. Infrastructure is sunk into and inside of other structures, social arrangements, and technologies. People do not necessarily distinguish the several coordinated aspects of infrastructure. In the Worm study, our respondents did not usually distinguish programs or subcomponents of the software—they were simply "in" it.

Transparency. Infrastructure is transparent to use, in the sense that it does not have to be reinvented each time or assembled for each task, but invisibly supports those tasks. For our respondents, the task of using ftp to download the system was new and thus difficult; for a computer scientist, this is an easy, routine task. Thus, the step of using ftp made the system less than transparent for the biologists, and thus much less usable.

Reach or scope. This may be either spatial or temporal—infrastructure has reach beyond a single event or one-site practice. One of the first things we did in system development was scan in the quarterly newsletter of the biologists so that one of the long-term rhythms of the community could be emulated online.

Learned as part of membership. The taken-for-grantedness of artifacts and organizational arrangements is a *sine qua non* of membership in a community of practice (Bowker & Star, in press; Lave & Wenger, 1991). Strangers and outsiders encounter infrastructure as a target object to be learned about. New participants acquire a naturalized familiarity with its objects, as they become members. Although many of the objects of biology were strange to us as ethnographers, and to the computer scientists, and we made a special effort to overcome this strangeness, it was easy to overlook other things that we had already naturalized, such as information retrieval practices over networked systems.

Links with conventions of practice. Infrastructure both shapes and is shaped by the conventions of a community of practice (e.g., the ways that cycles of day-night work are affected by and affect electrical power rates and needs). Generations of typists have learned the QWERTY keyboard; its limitations are inherited by the computer keyboard and thence by the design of today's computer furniture (Becker, 1982). The practices of reporting quarterly via the newsletter could not be changed in the biologists' system—when we suggested continual update, it was soundly rejected as interfering with important conventions of practice.

Embodiment of standards. Modified by scope and often by conflicting conventions, infrastructure takes on transparency by plugging into other infrastructures and tools in a standardized fashion. Our system embodied many standards used in the biological and academic community such as the names and maps for genetic strains, and photographs of relevant parts of the organism. But other standards escaped us at first, such as the use of specific programs for producing photographs on the Macintosh.

Built on an installed base. Infrastructure does not grow *de novo*; it wrestles with the inertia of the installed base and inherits strengths and limitations from that base. Optical fibers run along old railroad lines; new systems are designed for backward compatibility, and failing to account for these constraints may be fatal or distorting to new development processes (Hanseth & Monteiro, 1996). We partially availed ourselves of this in activities such as scanning in the newsletter and providing a searchable archive; but our failure to understand the extent of the Macintosh entrenchment in the community proved expensive.

Becomes visible upon breakdown. The normally invisible quality of working infrastructure becomes visible when it breaks: the server is down, the bridge washes out, there is a power blackout. Even when there are back-up mechanisms or procedures, their existence further highlights the now-visible infrastructure. One of the flags for our understanding of the importance of infrastructure came with field visits to check the system usability. Respondents would say prior to the visit that they were using the system with no problems—during the site visit, they were unable even to tell us where the system was on their local machines. This breakdown became the basis for a much more detailed understanding of the relational nature of infrastructure.

Is fixed in modular increments, not all at once or globally. Because infrastructure is big, layered, and complex, and because it means different things locally, it is never changed from above. Changes take time and negotiation, and adjustment with other aspects of the systems are involved.[1] Nobody is really in charge of infrastructure. When in the field, we would attempt to get systems up and running for respondents, and our attempts were often stymied by the myriad of ways in which lab computing was inveigled in local campus or hospital computing efforts, and in legacy systems. There simply was no magic wand to be waved over the development effort.

INFRASTRUCTURE AND METHODS

The methodological implications of this relational approach to infrastructure are considerable. Sites to examine then include decisions about encoding and standardizing, tinkering and tailoring activities (see, e.g., Gasser, 1986; Trigg & Bødker, 1994), and the observation and deconstruction of decisions carried into infrastructural forms (Bowker & Star, in press). The fieldwork in this case transmogrifies to a combination of historical and literary analysis, traditional tools like interviews and observations, systems analysis, and usability studies. For example, in studying the development of categories as part of information infrastructure, I observed meetings of nurses striving to categorize their own work

(Bowker, Timmermans, & Star, 1995), studied the archives of meetings at the World Health Organization and its predecessors arguing about establishing and refining categories used on death certificates, and read old newspapers and law books recording cases of racial recategorization under apartheid in South Africa (Bowker & Star, in press). In each case, I brought an ethnographic sensibility to the data collection and analysis: an idea that people make meanings based on their circumstances, and that these meanings would be inscribed into their judgments about the built information environment.

I have also worked with computer scientists designing complex information systems. I began this work as a kind of informant about social organization. At first, the computer scientists sought examples of real organizational problem solving in order to model large-scale artificial intelligence systems. They identified problems from the realm of complex system development, and asked me to investigate their analog in organizational settings, primarily of scientists and engineers (Hewitt, 1986; Star, 1989). For example, when designers tried to model how a smart system would determine closure for a complex problem, I investigated how this was managed in 19th-century England by a group of neurophysiologists debating the functions of the brain (Star, 1989), and made formal models of the processes that were fed back to the computer scientists.

This early work began in the 1980s, before the current development in information systems partnering ethnographers with computer scientists for the purpose of improving usability (as, for example, in the Worm Community Study). During the last decade, some ethnographers have created durable partnerships with system developers in many countries, especially in the areas of computer-supported cooperative work (CSCW) and human-computer interaction (Bowker, Star, Turner, & Gasser, 1997). This work has emerged from a number of intellectual traditions, including ethnomethodology, symbolic interactionism, labor process research, and activity theory (cultural-historical psychology), among others.

All of us doing this work have begun to wrestle with questions of scalability that inherently touch on questions of infrastructure. It is possible (sort of) to maintain a traditional ethnographic research project when the setting involves one group of people and a small number of computer terminals. However, many settings involving computer design and use no longer fit this model. Groups are distributed geographically and temporally, and may involve hundreds of people and terminals. There have always been inherent scale limits on ethnography, by definition. The labor-intensive and analysis-intensive craft of qualitative research, combined with a historical emphasis on single investigator studies, has never lent itself to ethnography of thousands.[2]

At the same time, ethnography is a tempting tool for analyzing online interaction. Its strength has been that it is capable of surfacing silenced voices, juggling disparate meanings, and understanding the gap between words and deeds. Ethnographers are trained to understand viewpoints, the definition of the situation. Intuitively, these seem like important strengths for understanding the enormous changes being wrought by information technology. The scale question remains a pressing and open one for methodological concerns in the study of infrastructure. It is an ironic and tempting moment—we have the promise of a complete transcript of interactions, almost ready-made "fieldnotes" in the form of transaction logs and archives of e-mail discussions. At the same time, reducing this volume of material to something both manageable and analytically interesting is a tough nut to crack, despite the emergence of increasingly sophisticated tools such as Atlas/ti for qualitative analysis. Yet, I know of no one who has analyzed transaction logs to their own satisfaction, never mind to a standard of ethnographic veridicality (see Spasser, 1998, for a good discussion of some of these problems).

And we are still stuck with the problem of where online interactions fit with people's lives and organizations off-line. In the Worm Community Study, I tried simply to scale up traditional fieldwork techniques—and I and my research partner ended up traveling to dozens of labs, doing entrée work for each one, interviewing more than a hundred biologists, and exhausting myself in the process. In the Illinois Digital Library Project, our social science evaluation team found that we had to transform our original study of "emergent community processes in the digital library" (via fieldwork and transaction logs) to a linked set of interviews with potential users and ethnographies of the design team while we waited for the system testbed to emerge, some 2 years behind schedule (Bishop et al., in press; Neumann & Star, 1996). We had to invent new ways of triangulating and bootstrapping along with the systems developers. These new ways of working broke old forms both for our respondents and for us.

TRICKS OF THE TRADE[3]

The following section examines several tricks I have developed in the previously mentioned studies, helpful for "reading" infrastructure and unfreezing some of its features.

IDENTIFYING MASTER NARRATIVES AND "OTHERS"

Many information systems employ what literary theorists would call a master narrative, or a single voice that does not problematize diversity. This voice speaks unconsciously from the presumed center of things. An example of this encoding into infrastructure would be a medical history form for women that encodes monogamous traditional heterosexuality as the only class of responses: blanks for "maiden name" and "husband's name," blanks for "form of birth control," but none for other sexual practices that may have medical consequences, and no place at all for partners other than husband to be called in a medical emergency. Latour (1996) discusses the narrative inscribed in the failed metro system, *Aramis*, as encoding a particular size of car based on the presumed nuclear family. Bandages or mastectomy prostheses labeled “flesh colored” that are closest to the color of White people’s skin are another kind of example.

Listening for the master narrative and identifying it as such means identifying first with that which has been made other, or unnamed. Some of the literary devices that represent master narratives include creating global actors, or turning a diverse set of activities and interests into one actor with a presumably monolithic agenda (“the United States stands for democracy”); personification, or making a set of actions into a single actor with volition (“science seeks a cure for cancer”); passive voice (“the data have revealed that”); and deletion of modalities. The latter has been well-described by sociologists of science—the process by which a scientific fact is gradually stripped of the circumstances of its development, and the attendant uncertainties, and becomes an unvarnished truth.

In the previously mentioned study of the International Classification of Diseases, Bowker and I discovered many moments when the master narrative-in-the-making became visible. One such deconstructive moment occurred when a committee of statisticians attempted to codify the “moment of life”: How can you tell, for the purposes of filling out a birth certificate, when a baby is alive? Religious differences (as, for example, between Catholics and Protestants) were argued about, as well as phenomenological distinctions such as the number of breaths a baby would draw, try to draw, or fail to draw (Bowker & Star, in press). In studies we read of the actual practices of filling in death certificates, the distinctions made by the “designers” upstream did not match the ways that attending doctors saw the world. We came to understand how the blanks on the forms were both heteropraxial (different practices according to region, local constraints, beliefs) and heteroglossial (inscribing different voices in the seemingly monotonous form).

SURFACING INVISIBLE WORK

Information systems encode and embed work in several ways. They may directly attempt to represent that work. They may sit in the middle of a work process like a rock in a stream, and require workarounds in order that interaction proceed around them. They also may leave gaps in work processes that require real-time adjustments, or *articulation* work, to complete the processes.

Finding the invisible work in information systems requires looking for these processes in the traces left behind by coders, designers, and users of systems (Star & Strauss, 1999, discuss this in relation to the design of CSCW systems). In some instances, this means going backstage, in Goffman’s (1959) terms, and recovering the mess obscured by the boring sameness of the information represented. It is often in such backstage work that important requirements are discovered. For example, in the Worm Community Study, we discovered that there were crucial moments in a biologist’s career—especially during the postdoc period, just before getting one’s own lab—where secrecy and professional smoothness are valued over the usual community norms of sharing preliminary results in semiformal venues.

With any form of work, there are always people whose work goes unnoticed or is not formally recognized (cleaners, janitors, maids, and often parents, for instance). Where the object of systems design is to support all work, leaving out what are locally perceived as "nonpeople" can mean a nonworking system. For example, with the biologists, I had originally wanted to include secretaries in the publication and communication stream, as they were so obviously (to me) part of the community communications. This was strongly resisted by both biologists and systems developers, as they did not see the secretaries as doing real science, and thus the idea was dropped. There is often a delicate balance of this sort between making things visible and leaving things tacit. With the nurses previously mentioned, whose work was categorizing all the tasks done by nurses, this was an important issue. Leave the work tacit, and it fades into the wallpaper (in one respondent's words, "we are thrown in with the price of the room"). Make it explicit, and it will become a target for hospital cost accounting. The job of the nursing classifiers was to balance someone in the middle, making their work just visible enough for legitimation, but maintaining an area of discretion. Without the fieldwork at their sessions where they were building the classification system, Bowker, Timmermans, and I (1995) would never have known about this conflict.

PARADOXES OF INFRASTRUCTURE

Why does the slightest small obstacle often present a barrier to the user of a computer system? One of the findings of our studies of users in the Illinois Digital Library Project (Bishop et al., in press) is that seemingly trivial alterations in routine, or demands for action, will act to prevent them from using the system. This can be an extra button to push, another link to follow to find help, or even looking up from the screen. The obduracy of these "tiny" barriers presents, at first glance, a puzzle in human irrationality. Why would someone not punch a couple of buttons rather than walk across campus to get a copy of something? Why do people persist in using less functional, but more routine actions when cheaper alternatives are nearby? Are people so routinized, so rigid in their ability to adapt to change that even such a slight impediment is too much?

Rather than characterize human nature with such broad strokes, I return to a fieldwork example for an explanation of this phenomenon. At a phenomenological level, what has happened is that these slight impediments have become magnified in the flow of the work process. An extra keyboard stroke might as well be an extra 10 pushups. What is going on here?

One way to explain this magnification process is to understand that in fact two processes of work are occurring simultaneously: Only one is visible to the traditional analysis of user-at-terminal or user-with-system. That is the one that concerns keystrokes and functionality. The other is the process of assemblage, the delicate, complex weaving together of desktop resources, organizational routines, running memory of complicated task queues (only a couple of which really concern the terminal or system), and all manner of articulation work performed invisibly by the user.

Schmidt and Simone (1996) show that production/coordination work and articulation work (the second set of invisible tasks previously described) are recursively related in the work situation. Only by describing *both* the production task and the hidden tasks of articulation, together and recursively, can we come up with a good analysis of why some systems work and others do not. The magnification we encountered in our studies of users concerns the disruption of the users' articulation work. This system is necessarily fragile (as it is in real time), depending on local and situated contingencies, and requires a great deal of street smarts to pull off. Small disruptions in the articulating processes may ramify throughout the workflow of the user, causing the seemingly small anomaly or extra gesture to have a far greater impact than a rational user-meets-terminal model would suggest.

THE THORNY PROBLEM OF INDICATORS

One of the difficulties in studying infrastructure is distinguishing different levels of reference in one's subject matter. This is a difficulty shared by all interpretive studies of media. For instance, suppose one wishes to understand the relationship of scientific advertising to cultural values about science. At one level of reference, one could count the frequency of ads, their claimed links with sales, and the attendant budget without even reading a single ad. In this case, the ads are indicators of resources spent promoting scientific products. Taking a step into the content of the advertisements, one could trace the emphases placed on certain types of activity, or the gender-stereotyped behavior embodied in them, or what sorts of images and aesthetics are used to display success. Here, one is required to assess the stylistics of the advertisements' creators—including ironic usage, multiple levels of meaning, psychological strategies employed, and thus their meanings. Finally, one could simply take the advertisements as a literal transcript about the process and progress of science, to be read directly for their claims, as indicators of scientific activity. To generalize this, one can read information infrastructure either as:

- a material *artifact* constructed by people, with physical properties and pragmatic properties in its effects on human organization. The truth status of the content of the information is not relevant in this perspective, only its impact; or as
- a *trace* or *record* of activities. Here, the information and its status become much more relevant, if the infrastructure itself becomes an information-collecting device. Transaction logs, e-mail records, as well as reading things like classification systems for evidence of cultural values, conflicts, or other decisions taken in construction fall into this category. The information infrastructure here sits (often

uneasily) somewhere between research assistant to the investigator and found cultural artifact. The information must still be analyzed, and placed in a larger framework of activities; or as

- a veridical representation of the world. Here, the information system is taken unproblematically as a mirror of actions in the world, and often tacitly, as a complete enough record of those actions. Where Usenet groups' interactions replace fieldnotes entirely in the analysis of a particular social world, for example, one has this sort of substitution.

These three sorts of representations are not mutually exclusive, of course. There is, however, an important methodological point to be made about where one's analysis is located. I have several times advised student theses that elide these functions of indicators, and it is a difficult and painful process to disentangle them. Films about rape may say a great deal about a given culture's acceptance of sexual violence, but they are not the same thing as police statistics about rape, nor the same as phenomenological investigations of the experience of being raped. Films are made by filmmakers who work within an industry, constrained by budgets, conventions, and their imaginations. Similarly, as an example from information infrastructure, people send e-mail according to certain conventions and within certain genres (Yates & Orlikowski, 1992). The relationship between e-mail and the larger sphere of lived activity cannot be presumed, but must be investigated.

The processes of discovering the status of indicators are complex. This is partly due to our own elisions as researchers, and partly due to sleights of hand undertaken by those creating them. A common example is the substitution of precision for validity in the creation of a system of indicators or categories. When large epistemological stakes are at issue in the development of a system, one political tactic is to focus away from the larger question, and instead to seize control of the indicators. Kirk and Kutchins (1992), in their study of the DSM, show precisely this set of tactics at work between psychoanalysts and biologically-oriented psychiatrists in the construction of that category system. Rather than (as they had in fact been doing for years) focus on the larger questions of mind and psychopathology, the designers of the DSM reframed the indicators, including how to frame requests for reimbursement from third parties, into a set of numbers that gradually squeezed out psychoanalytic approaches. I noted a similar set of activities by brain researchers at the turn of the century (Star, 1989).

BRIDGES AND BARRIERS

At least since Winner's (1986) classic chapter, "Do Artifacts Have Politics?" the question of whether and how values are inscribed into technical systems has been a live one in the communities studying technology and its design. Winner used the example of Robert Moses, a city planner in New York, who made a behind-the-scenes policy decision to make the automobile bridges over the Grand Central Parkway low in height. The reason? The bridges would then be too low for public transport—buses—to pass under them. The result? Poor people would be effectively barred from the richer Long Island suburbs, not by policy, but by design.

Whether or not one takes the Moses example at face value (and it has been a controversial one), the example is an instructive one. There are millions of tiny bridges built into large-scale information infrastructures, and millions of (literal and metaphoric) public buses that cannot pass through them. The example of computers given to inner-city schools and the developing world is an infamous one. The computers may work fine, but the electricity is dirty or lacking. Old floppy disks do not fit new drives, and new disks are expensive. Local phone calls are not always free. New browsers are faster, but more memory hungry. And one of those now popular will not support the most popular Web browser for blind people in text-only format.

In information infrastructure, every conceivable form of variation in practice, culture, and norm is inscribed at the deepest levels of design. Some are malleable, changeable, and programmable—if you have the knowledge, time, and other resources to do so. Others—such as a fixed-choice category set—present barriers to users that may only be changed by a full-scale social movement. Consider the example of choice of race in the U.S. Census forms. In the year 2000, for the first time, people may check more than one racial category. This simple infrastructural change took a march on Washington, years of political activism, and will cost billions of dollars. It is opposed by many progressive social justice groups, on the grounds that although it is biologically correct to say that most of us are multiracial, the effects of discrimination will be lost in the count by those who claim multiple racial origins.

Applying the insights, methods, and perspectives of ethnography to this class of issues is a terrifying and delightful challenge for what some would call the information age. The effort to date has linked historians, sociologists, anthropologists, philosophers, literary theorists, and computer scientists. The methodological side of the questions posed is underdeveloped by contrast with the power of the findings of this "invisible college." Thus, the articles in this issue are a most welcome addition to a literature of growing importance.

NOTES

[1] I am grateful to Kevin Powell for this point. This modularity is formally similar to Hewitt's (1986) open systems properties (see also Star, 1989).

[2] At least, that is, when those thousands are heterogeneous, distributed over many sites, and perhaps anonymous. Becker (personal communication, February 25, 1999) points out that some ethnographies of thousands have been done in large organizations (see, e.g., Becker, Geer, & Hughes, 1968).

3. This title is stolen from Becker's (1998) invaluable *Tricks of the Trade*, a handbook for conducting good social science research. The stealing, of course, is one of the key tricks of the trade. To quote Latour (1987), "les deux mamelles de la science sont peage et bricolage" (the twin teats of science are petty theft and bricolage).

REFERENCES

Bateson, G. (1978). *Steps to an ecology of mind.* New York: Ballantine.

Becker, H. S. (1982). *Art worlds.* Berkeley: University of California Press.

Becker, H. S. (1998). *Tricks of the trade: How to think about your research while you're doing it.* Chicago: University of Chicago Press.

Becker, H. S., Geer, B., & Hughes, E. C. (1968). *Making the grade: The academic side of college life.* New York: John Wiley.

Bowker, G. (1994). Information mythology and infrastructure. In L. Bud-Frierman (Ed.), *Information acumen: The understanding and use of knowledge in modern business* (pp. 231-247). London: Routledge.

Bowker, G., & Star, S. L. (in press). *Sorting things out: Classification and its consequences.* Cambridge, MA: MIT Press.

Bowker, G., Star, S. L., Turner, W., & Gasser, L. (Eds.). (1997). *Social science, information systems and cooperative work: Beyond the great divide.* Hillsdale, NJ: Lawrence Erlbaum.

Bowker, G., Timmermans, S., & Star, S. L. (1995). Infrastructure and organizational transformation: Classifying nurses' work. In W. Orlikowski, G. Walsham, M. Jones, & J. DeGross (Eds.), *Information technology and changes in organizational work* (pp. 344-370). London: Chapman and Hall.

Bucciarelli, L. L. (1994). *Designing engineers.* Cambridge, MA: MIT Press.

Clarke, A. E., & Fujimura, J. H. (Eds.). (1992). *The right tools for the job: At-work in twentieth-century life sciences.* Princeton, NJ: Princeton University Press.

Edwards, P. N. (1996). *The closed world: Computers and the politics of discourse in cold war America.* Cambridge, MA: MIT Press.

Gasser, L. (1986). The integration of computing and routine work. *ACM Transactions on Office Information Systems*, *4*, 205-225.

Goffman, E. (1959). *The presentation of self in everyday life.* Garden City, NY: Doubleday.

Goguen, J. (1997). Requirements engineering as the reconciliation of technical and social issues. In M. Jirotka & J. Goguen (Eds.), *Requirements engineering: Social and technical issues* (pp. 27-56). New York: Academic Press.

Hanseth, O., & Monteiro, E. (1996). Inscribing behavior in information infrastructure standards. *Accounting, Management & Information Technology*, *7*, 183-211.

Hanseth, O., Monteiro, E., & Hatling, M. (1996). Developing information infrastructure: The tension between standardization and flexibility. *Science, Technology and Human Values*, *21*, 407-426.

Hewitt, C. (1986). Offices are open systems. *ACM Transactions on Office Information Systems*, *4*, 271-287.

Hughes, T. P. (1983). *Networks of power: Electrification in Western society, 1880-1930.* Baltimore: Johns Hopkins University Press.

Hughes, T. P. (1989). The evolution of large technological systems. In W. E. Bijker, T. P. Hughes, & T. Pinch (Eds.), *The social construction of technological systems* (pp. 51-82). Cambridge, MA: MIT Press.

Jewett, T., & Kling, R. (1991). The dynamics of computerization in a social science research team: A case study of infrastructure, strategies, and skills. *Social Science Computer Review*, *9*, 246-275.

Kirk, S. A., & Kutchins, H. (1992). *The selling of the DSM: The rhetoric of science in psychiatry.* New York: Aldine de Gruyter.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society.* Milton Keynes, UK: Open University Press.

Latour, B. (1996). *Aramis, or the love of technology.* Cambridge, MA: Harvard University Press.

Latour, B., & Hermant, E. (1998). *Paris: Ville invisible.* Paris: La Decouverte.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press.

Neumann, L., & Star, S. L. (1996). Making infrastructure: The dream of a common language. In J. Blomberg, F. Kensing, & E. Dykstra-Erickson (Eds.), *Proceedings of the PDC '96* (pp. 231-240). Palo Alto, CA: Computer Professionals for Social Responsibility.

Schatz, B. (1991). Building an electronic community system. *Journal of Management Information Systems, 8*, 87-107.

Schmidt, K., & Simone, C. (1996). Coordination mechanisms: Towards a conceptual foundation of CSCW systems design. *Computer Supported Cooperative Work (CSCW): The Journal of Collaborative Computing, 5*, 155-200.

Star, S. L. (1989). *Regions of the mind: Brain research and the quest for scientific certainty.* Stanford, CA: Stanford University Press.

Star, S. L. (1991). Power, technologies and the phenomenology of conventions: On being allergic to onions. In J. Law (Ed.), *A sociology of monsters: Essays on power, technology and domination* (pp. 26-56). London: Routledge.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research, 7*(1), 111-134.

Star, S. L., & Strauss, A. L. (1999). Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer-Supported Cooperative Work (CSCW): The Journal of Collaborative Computing, 8*, 9-30.

Strauss, A. (Ed.). (1979). *Where medicine fails.* New Brunswick, NJ: Transaction Books.

Suchman, L., & Trigg, R. (1991). Understanding practice: Video as a medium for reflection and design. In J. Greenbaum & M. Kyng (Eds.), *Design at work* (pp. 65-89). London: Lawrence Erlbaum.

Summerton, J. (Ed.). (1994). *Changing large technical systems.* Boulder, CO: Westview.

Trigg, R., & Bødker, S. (1994). From implementation to design: Tailoring and the emergence of systematization in CSCW. In *Proceedings of ACM 1994 Conference on Computer-Supported Cooperative Work* (pp. 45-54). New York: ACM Press.

Winner, L. (1986). Do artifacts have politics? In J. Wacjman & D. Mackenzie (Eds.), *The social shaping of technology: How the refrigerator got its hum* (pp. 26-37). Milton Keynes, UK: Open University Press.

Yates, J. (1989). *Control through communication: The rise of system in American management.* Baltimore: Johns Hopkins University Press.

Yates, J., & Orlikowski, W. J. (1992). Genes of organizational communication: A structurational approach to studying communication and media. *Academy of Management Review, 17*, 299-326.

2 COMMENTS

Linnet Taylor – Towards a contextual and inclusive data studies: a response to Dalton and Thatcher

The social sciences are engaged in a trans-disciplinary debate on the meaning and use of new forms of digital data. One of the most important repercussions from Dalton and Thatcher's call (2014) for a critical data studies has been an awareness that researchers need to continually sensitise themselves to the contextualities of data's production and use (Kitchin, 2014; Graham and Shelton, 2013; Nissenbaum, 2010). This short essay responds to this ongoing debate, laying out the case for such an awareness and asking how we might better operationalise it in data studies. If researchers working with the new data sources – and geographers in particular – can learn to think across contexts in a more inclusive way, it may take us further toward realising big data's promise as a tool for social scientific research.

Like Dalton and Thatcher, I use the terminology of 'big data' as central to the process of imagining a more contextually aware data studies, since it is precisely because of 'bigness' that context tends to disappear. 'Big' can easily become a synonym for 'universal' in ways that can be both unreflexive and insidious. For instance, a focus on the analytical challenges of large and complex datsets tends to crowd out a more inclusive perspective in favour of a focus on the most active online population – the US – because it provides the greatest breadth of data. 'Big' is powerful, it is epistemologically deterministic (Cherlet, 2013), and it suggests a *truthiness* that gets in the way of reflexivity.

The power and traction that big data's truthiness currently enjoys – the idea of its universality and cultural flatness – tells us something about our own academic context. We are operating in a time of economic austerity that is decimating the capacity of the public sector to collect and act on its own data, and of geopolitical instabilities that are generating a desire for clarity and operationalisable research. Both these factors have played a role in the tremendous discursive power of big data in social scientific research and governance. It is supposed to produce solutions for every problem, despite our currently imperfect understanding of its risks and biases, and it is seen as essential to economic recovery and the creation of opportunity. There is even research funding being directed towards instrumenting people to perceive it more positively (EScience Center, 2015).

We, as researchers, inevitably play a role in this instrumentation, either proactive or resistant. The current call for reflexivity in critical geography, in particular, is a response to this involvement. Can we get beyond these pressures? Can we realistically engage with the global scale on which digital data are produced, the diversity inherent in their production, and the ways in which that diversity is in turn processed out of sight? Possibly not. Puschmann and Burgess (2014) in their study of the metaphors of big data show that it is perceived as something wild and non-human, despite the fact that all digital data is produced in socially mediated ways. The commonly used terminology of 'data in the wild' is a convenient fiction because it deemphasises this social mediation and absolves the researcher from the unweildy process of understanding the unfamiliar languages, cultures and institutional and political landscapes in which much data is generated. Big data analytical processes contribute to a sense that context is too big a problem to tackle, particularly since merging and linking datasets often creates exponentially more contexts to take into account. So how can the contextual be accessed and included in accounts of how big data is operating? And how can data's diversity be understood on a more global and inclusive scale?

One step towards answering these questions is to become more conscious of the radical asymmetries of power and technology that shape big data's production. Dalton and Thatcher recommend that researchers pay attention to the differential power geometries highlighted by Massey (1993), but data studies presents us with layered power geometries of both activity and data produced from that activity. This makes it necessary to examine the unevenness in the way that born-digital data are produced, collected and manipulated. Mark Graham (2015) in particular has called attention to the asymmetric ways that digital data represents those in lower-income countries and the global South, full of gaps, unknown spaces and biases that are hard to measure. A micro-level analysis of how connectivity has been becoming available to lower-income and marginalised groups (Taylor, 2015) demonstrates that access to the kinds of technologies that generate data as a by-product (primarily mobile phones and the internet) is highly uneven and interrupted.

This unevenness in data production suggests that big data's universality is at best a methodologically necessary illusion supported by publication bias: high-profile journals are keen to publish innovative big data analytics but do not demand that researchers are specific about the shortcomings of their data. In fact, knowing the shortcomings of one's data is also a challenge, since there is little research that explains what is missing. In particular, now that more than half of mobile phones are owned by the global non-elite (ITU, 2013) it is easy to confuse globally available data with globally representative data. People in lower-income places tend to produce sparser and less granular data because they have access to previous-generation devices. Further, fewer types of survey data are available on those areas, making it harder to gauge the validity of what is available. This means that data about lower-income places (i.e. most of the world) cannot tell us as much as data about higher-income places, yet sweeping claims are being made for it in terms of transforming human life and opportunities. These should be examined.

The patchiness of big data is also related to who controls its production. Power over big data analytics is oligarchic, at least where those data arise as a by-product of corporate-mediated processes such as communication, internet use or the use of sensors. Just as Morozov (2013) has warned that when we reify 'the internet' we are in danger of empowering certain interests over others, similarly, reifying big data as the 'god's eye view' (Pentland, 2011) may also risk handing over the power to understand data to the private sector interests who control much of the access and analysis. For example, corporate power often modulates the way that people become data producers through practices such as zero-basing, where new users in developing countries are offered mobile internet in a monopolistic model (e.g. Facebook's internet.org) including only a few 'partner' web services, limiting and skewing the signals they emit when they are online.

These asymmetries make it important to acknowledge the power politics that determine which data we get to see, and which remains uncollected, unanalysed or otherwise inaccessible. Although data production is global, the power to use data has much more of a core-periphery dynamic since most people worldwide do not have the chance to manipulate, channel or analyse data about themselves or their communities. Rather than gaining agency as conscious volunteers of data, the majority are instead becoming subjects of 'invisible systems' (Bowker and Star, 2000: 33) where technology firms and governments merely appropriate their data doubles for economic and political control. As Mann (2015) has pointed out, big data does not necessarily represent or empower just by existing. Instead, it gains representative weight where people can gain control over the signals they are emitting and transform that control into economic and political leverage.

If researchers can gain a clearer idea of these particular who's and what's of big data, we may better understand how to use it. Perceptions of what data is for and what may be done with it differ radically depending on one's location, because so too do concepts that are taken for granted by social scientists as semantically stable, such as open data, privacy, volunteered data, and even the internet (as with the example of zero-basing, where 'the internet' differs by device).

This diverse view of data's origins, meaning and use makes a strong case for interdisciplinary and transdisciplinary research. Instead, research on big data is subject to a strong pressure from funders to become extra-disciplinary by collaborating more with enterprise and helping to generate innovation. This suggests that data studies is at risk of orienting itself towards complementing this kind of marketable research, for instance by filling the gaps that such research tends to leave open around privacy and ethics. One role for critical research on data, then, is to de-instrument people and sensitise them to the diverse contexts of data's use and production. In contrast, a lack of attention to this diversity makes it possible to flatten out data's difficult unevenness, and inevitably diverts attention from the way data may serve certain populations at the expense of others, or channel resources to some places at the expense of others. For a data studies to be critical, it also needs to become more global. To do this, we must learn from those who are mapping these new colonial landscapes, and start to rise to the challenge of finding a more global perspective on the meaning and uses of data.

***Linnet Taylor** (http://www.uva.nl/over-de-uva/organisatie/medewerkers/content/t/a/l.e.m.taylor/l.e.m.taylor.html) is affiliated with the University of Amsterdam, International Development Studies. This commentary is based on her current research project, which looks at the asymmetries in data production, control and use between north and south, and the ways in which research ethics need to respond. See her article in* Environment and Planning D: Society and Space, ' (http://epd.sagepub.com/content/early/2015/10/06/0263775815608851.abstract)*No place to hide? The ethics and analytics of tracking mobility using mobile phone data.' (http://epd.sagepub.com/content/early/2015/10/06/0263775815608851.abstract) She can be contacted at l.e.m.taylor@uva.nl.*

Reference

Bowker G and S Star 2000 *Sorting Things Out. Classification and its consequences*. Cambridge, Massachusetts: The MIT Press.

Cherlet J 2014 Epistemic and technological determinism in development aid. *Science, Technology & Human Values, 39*(6): 773-794.

Dalton C and J Thatcher 2014 What does a critical data studies look like and why do we care? (https://societyandspace.com/material/commentaries/craig-dalton-and-jim-thatcher-what-does-a-critical-data-studies-look-like-and-why-do-we-care-seven-points-for-a-critical-approach-to-big-data/) *Society and Space Open Site*, May 20 2014. Accessed 22.4.15.

EScience Center 2015 eScience Integrator receives NWO funds for research on the perception of big data. Accessed 30.4.15. Available at https://www.esciencecenter.nl/news/escience-integrator-receives-nwo-funds-for-research-on-the-perception-of-bi (https://www.esciencecenter.nl/news/escience-integrator-receives-nwo-funds-for-research-on-the-perception-of-bi)

Graham M and T Shelton 2013 Geography and the future of big data, big data and the future of geography. *Dialogues in Human Geography 3*(3): 255-261.

Graham M 2015 The hidden biases of Geodata. Guardian datablog, April 30 2015. Available at: http://www.theguardian.com/news/datablog/2015/apr/28/the-hidden-biases-of-geodata (http://www.theguardian.com/news/datablog/2015/apr/28/the-hidden-biases-of-geodata)

ITU 2013 The world in 2013. Geneva: International Telecommunications Union.

Kitchin R 2014 Big Data, new epistemologies and paradigm shifts. *Big Data & Society 1*(1), 2053951714528481.

Mann L 2015 Left to Other Peoples' Devices: A Political Economy Approach to the Big Data Revolution in Development. Africa Summit, London School of Economics, 17th of April, 2015.

Massey D 1993 Power-Geometry and a Progressive Sense of Place. In: J Bird, B Curtis, T Putnam, G Robertson, and L Tickner (eds) *Mapping the Futures: Local cultures, global change*. New York: Routledge, pp 59-69.

Morozov E 2013 *To Save Everything, Click Here: Technology, solutionism, and the urge to fix problems that don't exist*. Penguin UK.

Nissenbaum H 2011 A contextual approach to privacy online. *Daedalus 140*(4): 32-48.

Pentland A 2011 Society's nervous system: building effective government, energy, and public health systems. *Pervasive and Mobile Computing 7*(6): 643-659.

Puschmann C and J Burgess 2014 Big Data, Big Questions | Metaphors of Big Data. *International Journal of Communication 8*: 1690-1709.

Taylor L 2015 Inside the Black Box of Internet Adoption: The Role of Migration and Networking in Internet Penetration in West Africa. *Policy & Internet* (ahead-of-print).

2 thoughts on "Linnet Taylor – Towards a contextual and inclusive data studies: a response to Dalton and Thatcher"