Fabian Neuhaus &
Timothy Webmoor

AGILE ETHICS FOR MASSIFIED RESEARCH AND VISUALIZATION

In this paper, the authors examine some of the implications of born-digital research environments by discussing the emergence of data mining and the analysis of social media platforms. With the rise of individual online activity in chat rooms, social networking sites and micro-blogging services, new repositories for social science research have become available in large quantities. Given the changes of scale that accompany such research, both in terms of data mining and the communication of results, the authors term this type of research 'massified research'. This article argues that while the private and commercial processing of these new massive data sets is far from unproblematic, the use by academic practitioners poses particular challenges with respect to established ethical protocols. These involve reconfigurations of the external relations between researchers and participants, as well as the internal relations that compose the identities of the participant, the researcher and that of the data. Consequently, massified research and its outputs operate in a grey area of undefined conduct with respect to these concerns. The authors work through the specific case study of using Twitter's public Application Programming Interface for research and visualization. To conclude, this article proposes some potential best practices to extend current procedures and guidelines for such massified research. Most importantly, the authors develop these under the banner of 'agile ethics'. The authors conclude by making the counterintuitive suggestion that researchers make themselves as vulnerable to potential data mining as the subjects who comprise their data sets: a parity of practice.

Keywords ethics; visualization; e-science; social media platforms; data intensity; Twitter

(*Received 28 April 2011; final version received 16 August 2011*)

Information, Communication & Society Vol. 15, No. 1, February 2012, pp. 43–65
ISSN 1369-118X print/ISSN 1468-4462 online
http://www.tandfonline.com http://dx.doi.org/10.1080/1369118X.2011.616519

Routledge
Taylor & Francis Group

Introduction

Researchers in the social sciences spend countless hours refining their designs, honing their methodologies and preparing for work in the 'field' in order to ensure the collection of useful data. The accelerated use of information and communication technologies has altered how data are gathered, stored, processed, disseminated and re-used (Hine 2006a, 2006b; Dutton & Meyer 2009). Within the United States and United Kingdom, programmes in cyber-science and e-science are promoted by governmental funding agencies such as the National Science Foundation, the Economic and Social Research Council (ESRC) and the Engineering and Physical Sciences Research Council. A component of this funding is to generate reflexive insights into changing practices in digitally enabled social research (Hine 2006a, 2006b, 2008).

In this paper, we examine some of the implications of born-digital research environments by discussing the emergence of data mining and the analysis of social media platforms. With the rise of individual online activity in chat rooms, social networking platforms and now micro-blogging services, new repositories for social science research have become available in large quantities (see Dutton & Jeffreys 2010 for an overview). Such 'network-enabled research', an infrastructural affordance for studying online sociality, is one facet of the development of a suite of 'e-research' or 'digital social research' strategies for understanding the consequences of the computerization of society (Hine 2006a, 2006b; Jeffreys 2010). While substantive impact on society based on technological change is rightly debated, the caveat being that supposed radical change may partly be 'cyberbole' linked to the marketing strategies within consumerist cultures (Woolgar 2002, pp. 9, 22), there remains little doubt that there has been an amplification of the number of individuals involved in the social analysis that deploys 'research-centred computational networks' (Dutton 2010, p. 21).

Embedded within the practices of digitally enabled research are ethical considerations (Carusi 2008; Carusi & Jirotka 2009; Ess 2009; Dutton & Piper 2010; Beaulieu & Estalella 2011). In offline settings, researchers wishing to collect or generate information relevant to social processes do so by enlisting the participation of individuals. Experimental psychology and social/cultural anthropology were early to recognize the need for codes of conduct for scholars enlisting participants in research in order that their well-being was considered.[1] Primary among the procedural mechanisms to protect individuals involved in research are privacy, confidentiality, anonymity and informed consent (Ess 2009; Dutton & Piper 2010, p. 224). Informed by subsequent national and international data-protection legislation,[2] the implementation and enforcement of professional codes of conduct mean that researchers must negotiate a series of overlapping legal, institutional and professional protocols (Carusi & Jirotka 2009, pp. 290–291). Institutional review board (IRB) procedures have

developed as a bureaucratic mechanism to insure compliance on the part of the researcher with ethical protocols. A focal point of these protocols is to reduce potential risk to the participants engaged in research projects. For example, the ESRC formulates potential risk to participants in research as follows:

> Risk is often defined by reference to the potential physical or psychological harm, discomfort or stress to human participants that a research project might generate . . . These include risk to a subject's personal social standing, privacy, personal values and beliefs.
>
> (ESRC 2008, p. 21)

Informed consent and anonymizing data are the primary technical and procedural steps for minimizing risks to human subjects in non-medical research contexts. These steps ensure that individuals are protected from direct identification and that they are aware of the potential uses of their contributions. In particular, to disclose to subjects the possibility of subsequent uses outside the research setting that might unintentionally or inadvertently reveal subjects' personal identities. While these protocols should indeed be integral to social researchers' considerations from the outset of planning and research design, many commentators have noted the growing gap between the requirements of these ethical guidelines and actual practices in digital social research (Robson & Robson 1999; Carusi 2008; Beaulieu 2010). While this cannot simply be ascribed to differences in research conducted with analogue as opposed to digital media, there are, nonetheless, medium-specific attributes that question the appropriateness of current ethical policies.

In this paper, we extend previous considerations and recommendations of ethical conduct in born-digital research environments by discussing the emergence of data mining and analysis of social media platforms (see Boyd & Ellison 2008 for an overview). First, we discuss the possibilities of online social networking data reservoirs. Then, we introduce an example that deploys the Twitter platform to visualize online activity across urban centres as virtual landscapes. We use the case study to highlight the ethical implications of working with this type of data for research in the social sciences. We conclude by translating principles of agile software development into 'agile ethics'. Rather than a defined set of codes or bureacratic machinery to ensure ethical conduct from the top-down, this mode of engagement has more in common with what has been described as an *in situ* creative and collaborative ethical practice that works bottom-up (cf. Allen 1996, pp. 176–177).

With the rise of individual online activity in chat rooms, social networking platforms and micro-blogging services, large quantities of data are being composed by social scientists for analysis. Non-academic research organizations and companies have been early to recognize the potential for studying social behaviour specific to the Internet (e.g. IBM Many Eyes, Microsoft Research, Xerox

Park and Yahoo! Research). The change in sample sizes, for instance, from 100 participants to 100,000, is a dramatic challenge in numerous ways, technically, politically, but also ethically. We argue that these new massive data sets, or what amounts to the emergence of the Internet as database, pose particular challenges when they are used by academic researchers. These challenges are made manifold by the augmentation of the capacity to distribute and access the results of such research, particularly in the form of web-based visualizations.

Given the changes of scale that accompany such research, both in terms of data mining and communication of results, we term this type of research 'massified research'. A primary challenge of such research is the question of ethical conduct on the part of the researcher. We suggest that the information generated by users of social media platforms and services cannot be considered equivalent to conventional types of offline information collected by social researchers. Current ethical protocols are not, therefore, adequate for the types of digital social research increasingly being conducted. To make our point, we discuss a particular case of academic research involving visualizations of Twitter's Application Programming Interface (API) feed, survey the overlap between academic and established commercial uses of this information economy and draw out the implications of the case study in terms of more generalizable alterations to the external relations between researcher and 'participant' and to the internal relations comprising the 'participant' and the data itself in massified research. We conclude by presenting some best practices for massified research and visualization utilizing such information. Foremost among these is the advocation of agile ethics. This is ethical conduct that is flexibly adapted to the research settings specific to online activity. Developing from this more general sensibility for online researchers, we make practical recommendations of tying such accessible and open data sets and visualizations to log files of metadata that are maintained at researchers' physical institutes (see Bowker 2000; Baker & Bowker 2007 on metadata and digital archives; and Hine 2006a, 2006b on databases and working practices). We also encourage the creation of statements of claimed responsibility by researchers as a measure to rethink data mining and assembly as authorship. This situates accountability at the local level. It also allows such log files and statements of responsibility to circulate as flexibly as digital data through attaching these to research contracts.

The new landscapes of massified research: the example of Twitter API feeds

As a reference, we would like to introduce the idea of data points as instances of personal information used in a digital context. When signing up for a new online service, for example, a varying number of data points are required to uniquely identify individuals, set up accounts and log in/out. In the simplest case, these

data points include name, last name, username, password and email. As most of us are familiar with, more data points are often requested from services in the form of address, house number, post code, city, country, phone number, secret question plus a secret answer, gender, birthday and so forth. Such data points are being sent back and forth between clients, service providers and third parties, in many cases in the background of our online activity. So with the possible exception of the username, they form pre-existing information. As soon as online activity begins, a cascade of data points begins to be collected. These traces or digital footprints are produced as part of the engagement with the service. Some of them are directly controllable by the user. Many of them, however, are simply too complicated to manage, not individually adjustable or preventable by the user. These types of online traces include technical details of connection and equipment; for instance, details of operating software, browser version, settings, language, locational information either through Internet Protocol (IP) and connection or from a specialized device and origin and destination. These technical details are supplemented by other browser activities such as clicks, interest, focus, topic, time stamps, preferences and so on. Every single activity is logged locally. It is also remotely recorded on distributed servers. Furthermore, in addition to these activity-based data points that are generated from moment to moment as we engage with websites, there are additional, long-term background elements, such as cookies, that store identity information and communicate them via the web (see Dutton & Piper 2010, p. 231 for webmetrics and ethics).

While it is not the intention to provide here the full extent of technical details of online communication (see, for example, Krishnamurthy & Wills 2008; Irani *et al.* 2009), an important point is that there are many more data points than what are visible on the surface of the activity – on our interfaces. All of these traces play a potential part in data mining. Provided digital social researchers understand how to access, compile and render such background information, any of these data points may be harvested for social science research. Indeed, the emergence of web-based methods and digital social research is due in large part to the recognition of such online information and the concomitant development of medium-specific tools to make it useful to researchers (e.g. Rogers 2004; Marres & Rogers 2008).

There already exist a vast range of approaches to harvesting and conducting research with the dispersed database, that is the Internet (e.g. Fielding *et al.* 2008; Dutton & Jeffreys 2010). As examples, such research might include analyses of Wikipedia editing practices (Viégas *et al.* 2004), the manner in which political controversies are manifested through the web (Marres & Rogers 2008; MACOSPOL 2011) or the application of social network analyses to Facebook (Hogan 2008). These types of analyses of online information are, particularly when they include the spatial information embedded in web infrastructure, often referred to as 'neo-', 'supra-' or 'zero-geography'. While the idea that data

intensity necessitates compression through the visual register is by no means limited to Internet research (Tufte 1983), we note that due to the general invisibility of much of our online activity combined with the tremendous amounts of data points involved, nearly all of these web-based research approaches foreground visualizations as outputs of their results.

We are going to examine a particular subset of such research: Twitter. This social networking and micro-blogging service was launched in 2006 and as of early 2011 counts nearly 200 million users worldwide (Chang 2011). A number of research projects have begun to utilize Twitter for research (e.g. Leavitt 2009; Huberman *et al.* 2010; Wu *et al.* 2011). For our purposes of highlighting ethical implications, we will discuss a project developed at the Centre for Advanced Spatial Analysis at University College London that collects data points from Twitter to conduct research and communicates the results through visualizations. In this particular work, large data sets pulled from Twitter have been mapped as 'virtual' city landscapes (Figures 1 and 2).

With each map of the New City Landscape (NCL; Figure 2), we are working with approximately 150,000 location-based Twitter messages sent by about 15,000 individual Twitter users. The data are collected through the public Twitter API which is freely provided by Twitter.[3] An API is an Application Programming Interface, allowing external development of functions by providing a communication standard. One service offers this standard for other services to plug in and communicate with the first one. In most cases, such as the Twitter API, this primarily involves the exchange of data and functions. A goal for providing the API is, of course, to increase Twitter usage through extensibility.

Put simply, Twitter is providing other machines access to a data stream bundled with tweets. Using this API, Twitter packages the outgoing data stream of tweets for third party developers of Twitter applications. This means access to much more information than there would be as a normal user at the 'front-end' or browser interface where access is limited to followers and the content of messages of those followed. We underscore that these data points are collected remotely through the Internet, without a direct consent from the 'users' of Twitter who are posting messages.

Using the API, we can harvest a set number of data points from Twitter (Figure 1). These data points include TwitterPost (actual message); TwitterID (in addition to the username every user is assigned an internal number); dateT (date and time when the message was sent); name (user or screen name); link (link to online location of this tweet); usage, Twittergeo, lang (language); profile (link to online user profile); google_location (location of user derived as a geocoded place via the user profile location); atom_content (message in atom format); source (platform used to send the tweet); Lat (latitude) and Lon (longitude).

With a spatial search, we can filter the messages according to a specific location. For the NCL maps, we have defined the location consistently as an area within a 30 km radius around an urban centre. This search query will pass

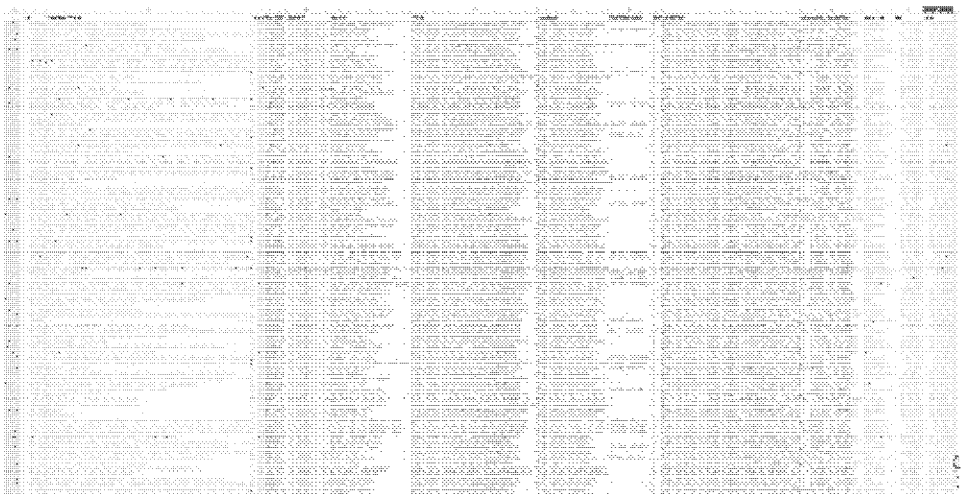

FIGURE 1 Image shows a typical raw data table between lines 33,470 and 33,558.

FIGURE 2 Two NCL maps as examples of aggregating potentially identifying information from Twitter. On the left is Moscow with a number of very active locations. Paris on the right exhibits a central main core of activity.

down from the Twitter feed all messages that fit this criterion. Via a special software, data are then stored continuously in a database.[4] All of this happens in real time.

Twitter is not actually sending all worldwide tweets through the Search API. What is sent is a random sample. This sample tends to be only around one per cent of all actual tweets. Nevertheless, given the nearly 200 million Twitter users sending an estimated 65 million tweets per day, this still provides a massive data

set (Chang 2011). A typical data table (Figure 1) may have a 150,000 tweets providing over 1.5 million discreet points of data. So, a data file for a week's worth of Twitter activity in London could on average contain between 100,000 and 150,000 lines and be upwards of 150 MB.[5] Given this scale, the data collection must be limited. So per geographical location, primarily international cities, collection has been limited to seven days of consecutive logging of messages sent using the Twitter service.

Already we see in this emerging context of research how the virtual and remote nature of data collection raises ethical questions. To what extent, for example, do the users of online services agree to 'their data' being used for further research or analysis – potentially useful information which they often unknowingly generate while online. Such data gathering already challenges the premise of informed consent for offline data collection and research. Futhermore, such data are often both personal, in terms of the content of the messages, and personally identifying given the background data points. Foremost among the concerns is, of course, the ease of access to locational information. The spatial information accompanying such messages begs the question of protection of privacy and undue surveillance. With the resulting data set containing the messages tagged with actual global positioning system (GPS) or latitude and longitude coordinates, it is theoretically possible to map an individual's location at the time of sending a message. Given the normal parameters of accuracy for GPS, an individual's location may be determined to within 5–15 metres. We will return to issues of spatial identification later. Now we continue with the subsequent processing and visualizing of such data to draw out these ethical implications.

Visualizing Twitter

For the NCL maps (Figure 2), this geographical information is used to produce a map showing the spatial dimensions of the data. For mapping, the individual points are being aggregated as a density surface (Quantdec 2003). With such a transformation, visualization provides detailed modelling of the landscape and enables a qualitative 'good reading' of the three-dimensionally interpreted landscape. The results are similar to topographic maps (Kraak 2010, p. 105). With this transformation, the map now communicates the active locations of Twitter usage via a three-dimensional landscape with high and low points, where mountains rise over active locations and cliffs drop down into calm valleys, flowing out to 'tweet deserts' where little to no tweets are sent.

The design of the maps is borrowed from classic hypsometric landscape maps. The colour scheme has been borrowed from a scheme developed by Rudolf Leuzinger (1826–1896) for the Carte physique et géographique de la France published in 1880 (Jenny & Raeber 2008). Each step-up represents a reduction of 10 per cent in the message group. That is, in the dark blue-green area, 100 per cent of

the logged messages are represented, whereas in the light-brown area of the top peak, only 10 per cent of the total number of messages recorded are represented.

The defining landscape features of the virtual NCL maps are the hot spots of Twitter activity – the peaks. Here, the morphology varies dramatically between the different observed urban areas. How Twitter traffic structures the NCL is unique to each city. There are, however, some common characteristics that can be pointed out.

The emerging landscape features have been renamed to reflect these new landscape conditions. New labels have been fabricated using real world names combined with descriptions of the virtual surface overlay. As such, it puts a stronger emphasis on how we talk about place and how it is picked up in everyday language (de Certeau 1984).

Given the ephemeral nature of the Twitter data, temporal aspects on the level of the urban scale are very much of interest. This is not directly apparent in the maps. Therefore, timeRose diagrams focusing on the temporal dimension accompany the maps (Figure 3).

The resulting visualizations of both the NCL maps and the timeRose diagrams involve individual tweets to render an overview of the aggregated data set. This has consequences for the ethical implications of such research. While the individual message is key to the visualization process, it does not feature in the final product. Through transforming the numerical and textual information of Twitter API into the visual register, personally identifiable information has been removed. This 'anonymization' results from the scale at which these visualizations work. We return to this point later in terms of the resolution or granularity of massified visualizations. Nevertheless, as discussed above, lots of data points, including potentially sensitive information, are part of the raw data sets and worked with during the assembly process. Ethical considerations are, therefore, wrapped up with more than the outputs of such visualizing research. As with other types of research, we must consider the entire production process, including gathering and storage as well as the transferring and re-use of such data sets.

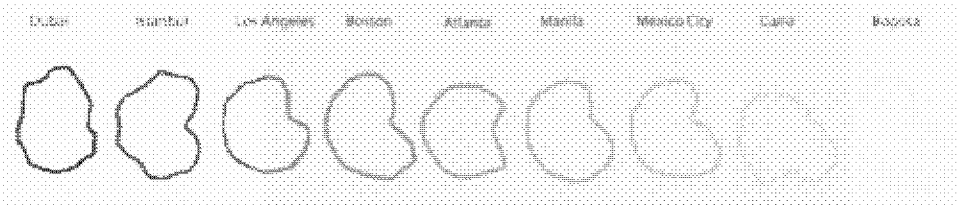

FIGURE 3 Ordering nine cities from around the world according to the Twitter activity by hour of the day. Far left represents evening, and far right is the morning. Cities in the centre show less preference for Twitter activity at certain times of the day. Each shape follows a 24-hour clock with midnight being on the top and midday at the bottom.

Re-mixing private–public domains: perceptions of online activity

With Twitter, the user can set up a personal profile and start sending 140-character messages. These messages are generally undirected statements that are sent out to the world using the Twitter platform. This is the micro-blogging aspect of Twitter. However, as with the blogosphere, there is simply too much – and ever increasing – 'noise' ready-to-hand on the Internet. To sort signal from this noise, Twitter works similarly to other social networking sites (SNSs). All possible tweets in the 'Twitterscape' are filtered through social connection. So to have other people's messages delivered onto the personal Twitter account page, one starts 'following' other users. This has to happen in order for other users to see one's messages. Each user can manually manage the list of followed accounts and followers.

However, while this setting to share tweets among a select group creates a sense of closed community and might lead one to believe that the information or data sent using this platform can only be read and accessed by the circle of followers (e.g. 'friends'), this is patently not the case (Boyd 2008; Boyd & Ellison 2008). Every Twitter message sent is public, unless specifically sent as a private message. We, therefore, define the terms private and public as used in the context of such social networking as follows: private being exclusive to a specific individual or group with managed access and public being acessible to any interested party with unmonitored access to services and information.

At the interface of our displays, this blurring of perceptions of private and public is not, however, unique to Twitter. It has, of course, been remarked upon more generally as a facet of sociality in the age of the Internet (Rheingold 2000; Boellstorff 2008; Boyd 2008). The screen or graphical user interface as the locus of underlying spatial and temporal transformations has been most forcefully theorized by Paul Virilio (Friedberg 2004). Indeed, the issue arises in a number of fields related to user-generated data, ranging from Google to Facebook, from Microsoft to Apple and from loyalty cards to travel cards. Information is the basic currency of the information society. User-generated data on the web are constantly being analysed and pored back into the ocean of data. While many online service providers initially reacted against being part of the information economy, citing values of digital democracy, free software and open source movement mottos, most have turned to monetizing their services (e.g. Rheingold 2000; Weber 2004; see Auletta 2009 on Google's AdSense campaign). Now, such data mining for commercial purposes is to large degree built into web platforms.

Given that the information the user generates on the Internet leaves traces by the click and beyond, there is increasing commercial incentive to harvest, archive and analyse our online 'digital heritage'. However, the traces we create are not

limited to the past (see Webmoor 2008; Harrison 2009 on 'archaeology' of the Internet). It travels beside the user in the present, arriving beforehand at the shores of potential service providers almost like a rippling wave in the ocean of the web. Tracking of online activities, in terms of content, activity, time and physical location, is increasingly possible. This is especially so with Twitter and similar location-based media platforms. Previously, projects wishing to collect precise physical locations of Internet users have had to rely primarily upon GPS technology or mobile phone data (for example, Eagle *et al.* 2009 and their study of social networks based on the use of mobile phones).

The implications in the case of Twitter, and other similar social media that provide this 'backend' of user profile information through 'dumps' or APIs, develop from the perceptions of private and public. In agreement with other scholars who are engaged in social research of web-based activities, the sliding scale of publicness and privacy poses particular problems for ensuring virtuous conduct on the part of the researcher (Frankel & Siang 1999; Bakardjieva & Feenberg 2001). Moreover, as is typical of online researchers, we also find ourselves in the grey area or slippages along this spectrum of activity. There is then a doubling of this distinction. This involves the generation of online traces or information by 'participants' in relation to perceptions of occupying a public or private setting. There is also the issue of researchers' access to such information and whether this is done publicly or privately and without public awareness. We suggest that while the former blurring has long been an issue, the latter has only more recently come to the forefront of attention (for example, virtual ethnographies and 'lurking' in online forums; see Bakardjieva & Feenberg 2001; Boellstorff 2008 for discussions).

As described above using the example of Twitter, the issue of privacy is not clear cut. This is because within the information economy, privacy is perceived by users in one manner, yet is handled by the online service provider in a very different manner. As an interesting point of comparison, let us consider commercial physical space. More and more public spaces are merging into corporate spaces in the city. Shopping malls, as concrete examples, start to enter the domain of the space perceived as 'public'. Even though these buildings are part of a privately owned mall, designed so that money is generated, they are successfully camouflaged as public space where people gather socially, recreate, entertain themselves and, of course, spend money. There is even a selection of shopping 'peers' through economic, locational and social targeting. Yet as private spaces, information collection and surveillance footage are being continuously and lawfully collected.

It could be argued that web services are quite similar to what is described above. We are not surfing the 'public' Internet as such, even though most websites are free to use and easily accessible. They are actually privately owned and most often offering a service. Of course, the service provider will expect financial gain. If not directly from the user, then through a third party that offers

money in exchange for another commodity or service; mostly, the directing of users to other sites and/or information such as advertisements. In this sense, the user is provided with a free service in exchange for letting himself/herself be directed to potentially interesting information and adverts. Despite the seeming reasonableness of this fair exchange principle of the information economy, the Internet remains especially ambiguous in terms of just what is being 'exchanged' by users.

Facebook has a number of webpages dedicated to the topic of privacy: one, for example, to explain the different settings categories[6] and another for the privacy policy.[7] The changes since the launch of Facebook in 2004 have always been loudly commented upon. Recently, Matt McKeon has put together a visualization of the evolution of Facebook privacy over the past five years (see http://mattmckeon.com/facebook-privacy).[8]

Like Facebook, Twitter also has a privacy page where they attempt to explain the company's privacy guidelines and considerations. As of March 2011, it states: 'We collect and use your information to provide our Services and improve them over time'.[9] On this webpage, Twitter clearly states that the concept of the service is to publicly distribute messages. 'What you say on Twitter may be viewed all around the world instantly'. It further states that the default setting is set to public with the option to make it more private. This is not true, however, for the locational information. Each user must choose to activate this feature of Twitter. There is also an option to opt out of this and retroactively delete the locational information of all messages sent in the past: 'You may delete all locational information from your past tweets. This may take up to thirty minutes'. In the strict sense, therefore, every user who's locational information is mapped onto the NCL maps has chosen to share this information with the world.

Resolution of ethics through resolution of images: anonymity through data deluge

Despite these attempts at disclosure by social media platforms as to how users' information may be publicly accessible, there remain important ethical considerations. The first consideration is just what is being made public through Facebook or Twitter's API. It is important to clarify that access to an individual's profile and digital traces extends to his/her network of online interactions. That is, the data of interest for a whole range of commercial and academic or political bodies are not confined to just the actual content of a Twitter message. As discussed earlier, each account or profile contains additional information, such as name, age, gender, address, contact details, interests, birthday, shoe size and so forth. All of which may be extremely valuable for purposes of market research, focused advertisement or compiling background demographics. The

big potential is, however, what has been alluded to with Facebook. Connections and social networks may be immediately available or at least reconstructed from online services. Who knows whom, who is contacting whom, when, how often and where. This is the real aspect of change with personal identify information. For the first time, we can actually observe large-scale social interaction in dramatic detail in real time and describe the individual in the context of activities, preferences and connections.

These connections or data pertaining to networks are also increasingly grafted onto offline, geographical locations. With the emergence of ubiquitous computing and mobile devices that integrate location-based software, services and GPS, the pressing ethical consideration becomes the public nature of an individual's physical location. As alluded to above in relation to perceptions of public and private space, issues of surveillance become involved with data mining. Now, almost all services integrate actual locational data, either by using the integrated GPS module if used on a smart phone or IP or Wi-Fi access point data. Service providers know not only with whom one is connected but also where one is actually located when using the service.

The most public discussion around this very issue was the recent controversy involving Google and its launch of its Google Latitude service. The Latitude service would offer the option to distribute one's location to a list of friends who could follow one's movement in real time. Though they did provide a Google Privacy Statement, the controversy arose over a very specific scenario that involved access to, and use of, an individual's location. As it was presented, there was the possibility that a jealous husband could potentially log into the service and activate the service on his wife's mobile without her knowledge and thereby obtain his wife's position in real time delivered onto his screen (Dunt 2009; Kiss 2009). While a somewhat fanciful scenario, such a possibility became realized by commercial services who sprang up to mine and offer such locational information of individuals. For instance, the 'creepy software' stalks individuals.[10] By entering a username, the program returns all of the information pertaining to location that it can collect through Twitter and Flickr. It then displays it on a map, unveiling an individual's travelling habits and favourite locations. In response, companies such as Google initiated a weekly email reminder for users who had activated their Latitude service.

The potential misuse and danger to individuals as a consequence of data mining online social media was most forcefully brought to pubic attention by the platform 'pleaserobme.com'.[11] This service displayed information collected from SNSs of people who posted messages stating that they were not at home. In highlighting such statements, the service intimated that it would be the opportune time to burgle these individuals' houses. As we have shown for Twitter, this was made possible through the message embedded locational information.

The implications of the detailed knowledge of private information, and especially locational data, are that the identification of individuals for third parties becomes possible and that this information may potentially be used to harm the individual. In addition to the concerns of location and social network, a final issue that must be raised is that of the spatial resolution of such publicly accessible information. Having the information is not the same as being able to use it. It is a question of accessing and making it available. Indeed, there might be a degree of anonymity in the fact that the data pool is so vast that an individual's personal information is actually no longer visible. This is game deciding when the actual output of such mined private information is rendered visually.

For example, let us return to our case study. While the NCL maps are based on individual Twitter messages, the data have been aggregated and the resulting visualization is a density surface generated from the tweets. Consequently, the individual tweet no longer features in the output. Even if, for example, we show the location of an individual message as in the visualizations from a map of London generated with Twitter usage (Figure 4), the resolution of the clip in pixels is so low that it becomes nearly impossible to determine a precise location. The blurred pixels display what is more correctly thought of as a potential area; that is, the approximate physical coordinates where the individual is located when they send a message. In addition, as previously mentioned, the

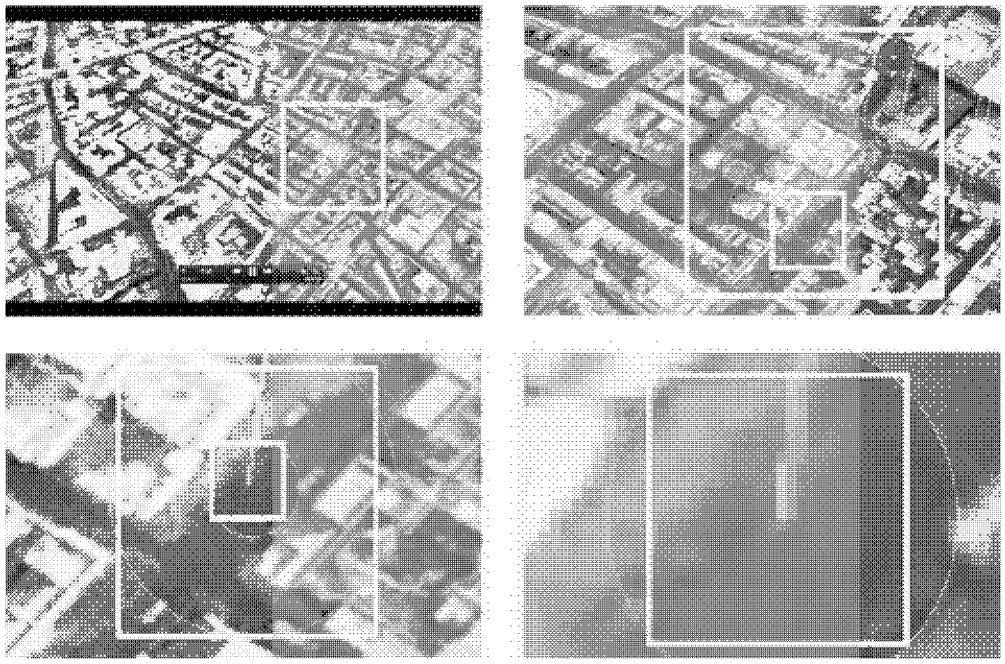

FIGURE 4 Images by urbanTick. This shows a zoom on an animation of tweets in Google Earth to demonstrate how granularity inhibits identifying an individual's precise location. The poor resolution is compounded when one factors in the limits of GPS accuracy.

accuracy of GPS is between 5 and 15 – at times 100 – metres in a dense urban environment. Combine this with the population density of London, for instance, and it becomes impossible to pinpoint an individual's precise whereabouts.

Ethical implications

Conventional guidelines for social science research are steered by IRB's human subjects protocols. These are, of course, fine as far as they go for 'offline research'. However, as we have attempted to illustrate, gaps widen not only because of scale, the sheer number that research with public API's such as Twitter's involves, but also because of the mixing and doubling of the private/public spectrum that is involved when we engage with such social media. There are numerous complications regarding the usage of these online data provided directly by a service such as Twitter. While there remains, to be sure, ethical issue at stake when such information is used by commercial parties, some of these uses may be predicated upon the exchange or transaction of mutual benefit between service providers and customers; in these cases, online SNS services. However, for academics in e-social science or digital social research who deploy such data, the ethical implications must be carefully evaluated. Especially so as the current IRB ethical protocols for academic researchers have, as we mentioned at the outset, yet to develop best practices that accommodate such emerging forms of massified research and visualization.

We identify five intertwined implications specific to such massified research that challenge conventional, offline research and the attendant ethical guidelines that have been established to minimize risk to the individual. Unfortunately, for issues of accountability, the combination of these changes to relations currently encourages the complete oversight of ethical measures. At root because there is easy access to the data that bypasses any 'producer' or subject.

Primary among these is the change in the enactment of the participant and researcher relationship. This fundamental change is due to the computer-mediated setting that characterizes the entire research process, from 'production' or harvesting, to processing and visualization of the data. As we have seen with Twitter's API, a researcher can conveniently collect data from his/her computer without the knowledge of the data producer or the 'participant'. Where IRB protocols could aid in maintaining a direct and formally recognized agreement between the two parties, now there is often no interaction between the two parties involved. We might say this is a change in the external relations of this type of research. These relations are now much more anonymous. In terms of accountability, the issue of (researcher) anonymity now becomes a problem rather than being a solution for minimizing risk to the individual in conventional research.

A second implication that likewise involves a change in the external relations has already been alluded to at the beginning of this paper. This is the sheer scale

of the research in terms of the number of 'participants' involved. Whether academic or commercial, researchers can mine the data points or digital heritage of 'participants' at an unprecedented scale. A different set of constraints comes into play in terms of determining a typical sample size for social scientific research. Whereas these constraints may have principally involved offline resources, such as personnel to distribute, collect or enter information from questionnaires, we now have the infrastructural constraints of server storage and API protocols. While these constraints still limit the number of individuals involved, the upper limit has been vastly augmented. Now it is possible to consider research conducted not with 1,000 or 10,000 participants but with 100,000 or one million individual users. The relations between researchers and participants are now one or a few to the multitudes. With large data sets, the individual disappears in the mass, but the issue and potential risk remain on the scale of the individual. Moreover, IRB protocol measures for informed consent are simply not viable at this scale.

Further implications to consider have to due with internal relations that comprise the identities of the 'participant', the researcher and that of the 'data'. What is the status of the 'participant'? Equally, what is the status of the 'data'? And under what conditions does the researcher engage with participants and data? Dissimilarly to offline settings, the identity and status of all of these are less clear cut and stable. As with the Twitter example, the internal relations that make apparent when an individual is engaged in private activities or a participant in public (or commercial) research are far from straightforward. While we have come to associate mixing and mashups with digital content and data, there is a sense in which the roles that an individual takes part in when using online social media are rapidly being mixed. Being a participant or a 'user' in research certainly involves the perception of whether one is engaged in public or private activities. But more frequently individuals are a mixture of the two, both using services and contributing to commercial and academic research. Offline notions of parsed and relatively stable identities and settings have been imported as parameters for online research. This is not always helpful. It obfuscates how our online status is an internal flux between these roles.

Likewise, the data collected often undergo re-purposing and remixing. Data are most often not produced with the intention that it be used as raw research material.[12] Additionally, the original data are, in the process of mining and visualizing, taken out of its context of production and assembled as a new instance. The data points made public while using online services are usually included in the raw data used in the research. If they are not, they, nonetheless, are simple to reconstruct as all instances are given an ID and uniquely linked to an account. While data points do not normally enable anyone in possession of the information to directly access a user's account, it can lead to secondary information. This is especially the case if traces are available over a longer period of time or in greater quantity so that a more comprehensive picture about an individual may

be pieced together. Even though most online presences of individual users are to some extent disconnected from the offline setting with, for example, the deployment of different screen names or invented bioinformation, individuals, nonetheless, tend to use the same identifiers across multiple accounts. Given this, a cross screening could identify the same individual through different online services.

Finally, where we might have several researchers and often a number of research assistants involved in the creation of a data set with a few hundred participants, now a single individual may establish the parameters and infrastructure for a more extensive social network data set containing five million individual users. In terms of accountability, this means that the context for peer feedback, monitoring and support is denuded. In consequence, the informal but pervasive measures for ensuring ethical conduct along the entire chain of research, from harvesting to processing and production of outputs, are attenuated. So instead of the risk of having the so-called information silo, we potentially have instead the information scientist silo.

Conclusions: agile research ethics

Given these changes in relations of accountability and identity with massified research, we suggest some best practices. First, it will be important for both researchers and institutions to accept the fact that this kind of large-scale data mining still involves human subjects. Individuals may indeed have agreed to the service conditions of a third party. However, this cancels neither the responsibilities of the researcher nor the institutions that undertake such research. As a practical measure, we suggest that institutions put a process in place for researchers to log any such data collections in advance. This can be used with the intention to monitor activities and to some extent bolster the attenuated peer supervision. On the other hand, it supports the researcher as a backup and support facility. It will encourage more rigorous planning and foster the careful treatment of the data in terms of use, storage, accessibility and publishing.

Second, as it will not be possible to put in place a contract between researchers and participants, given the scale and features of data collection, we need to place data generation on more of an equal footing with final outputs; to think of it in terms of authorship. Therefore, a secondary justification, in the form of a commitment or statement by the researcher, could be a tool to ensure quality and responsible handling. This would both break the spell of anonymity currently surrounding these data sets and assign a claimed responsibility. Given the programming skill and creativity involved in mining data, the data sets themselves, not solely the visualized outputs, become a contribution that researchers will want to take responsibility for.

Approaching mined data with such a personal and engaged attitude will be critical to fostering several on-the-ground practices. However, beyond these

practical considerations, we believe that developing an agile ethics will be most effective in avoiding the most pressing concern; that is, the current ability to consider mined data as disconnected from ethics. This is because, at root, the gaps in IRB protocols when applied to such massified research cannot be filled. For example, with the type of research we conducted using Twitter, there is no practical manner of informing several tens of thousands of individuals. Indeed, such informed consent is not even, at present, required by our IRB. More to the point, we feel that the development of bureacratic procedures for this type of online research will be ineffective or impractical if they are based on importing offline precedence. As opposed to adopting forms of situationist ethics or deontological ethics, we need medium-specific ethical conduct. This is why instead we advocate an agile ethics, the principles of which follow from those of the agile software development community. In particular, the agile principles place emphasis upon individuals and interactions over processes and tools, working solutions over comprehensive documentation, collaboration over contract negotiation and responsiveness to change over procedure (Shore & Warden 2008). This is more an attitude, or a mode of engagement and sensibility for good practice, as opposed to a formal list of procedures and protocols (cf. Allen's 1996 'dialogic' ethics). An agile research approach is form fit for the broad programme of massified research and visualization in the e-sciences. It also adjusts to the specific settings of analysis. Such flexibility and practice-related contingency are integral to agile research.

Furthermore, agile research enfolds ethical conduct in several ways that we feel is appropriate to the challenges identified for massified research. As an agile researcher, we become part of the network we work within. Our network 'contracts' expand our relations of accountability beyond the funding agencies, peers and IRBs to include the individuals/collaborators, institutions and technologies we research with. A primary facet of conduct with agile ethics is vulnerability. Against established IRB protocols that attempt to ensure some measures of privacy for the researcher, we advocate the counterintuitive practice of making ourselves vulnerable. Or, put another way, there should be publicity commensurate to research, or a 'parity of practice'. As we have discussed, the public and private realm are mixing with digital research, particularly research concerned with the Internet as database. As never before, the anonymity of ourselves as researchers and the individuals and settings we work with are difficult to maintain. Most protocols and perspectives discuss how to raise firewalls and maintain discretion despite increasing connectivity. An agile ethics makes the counterintuitive move to increased openness and transparency; to expose ourselves equally with those wrapped up in our projects. If we generate, study or deploy potentially personal information in our research, then our level of privacy ought to match that of the individuals involved in the project. Thus, as with our example of Twitter's API, these data are only provided if we do not keep our tweets private and if we opt for disclosing our location. If we as

researchers are involved with harvesting such information for academic outputs, we ought to allow our own data to be collected. That is, our vulnerability ought to be commensurate with that of the research 'participants'. We ought to go 'public' so that other researchers may potentially mine our own data points streaming through social media APIs.

Of course, such parity of practice will vary according to specific platforms deployed in research. But such agility and sensitivity to our socio-technical network are precisely the points with agile research ethics. Rather than feign that we as researchers may draw back to the shore, we recognize that we are already collectively immersed in the ocean of information that is the Internet.

Acknowledgements

An earlier version of this paper was presented at the Visualisation in the Age of Computerisation conference held in March 2011 at the University of Oxford. The authors would like to thank the participants for their helpful commentary, as well as the four anonymous reviewers and the editors for their insightful suggestions. The Twitter data collection for the NCL maps was initiated as a collaboration between Dr Andrew Hudson-Smith, Steven Gray and Fabian Neuhaus. The code to collect the Twitter data was developed by Steven Gray as part of the National e-Infrastructure for Social Simulation project. Support for this study was provided by the ESRC Oxford e-Social Science project.

Notes

1 For example, see the Code of Ethics of the American Anthropological Association, http://www.aaanet.org/committees/ethics/ethcode.htm, or the American Psychological Association's Code of Conduct, http://www.apa.org/ethics/code/index.aspx.

2 For instance, the UK's Data Protection Act of 1988 or the 1995 European Union Data Protection Directive.

3 See http://apiwiki.twitter.com/w/page/22554648/FrontPage for details.

4 Due to IP limitations imposed by Twitter and infrastructural limitations, only four parallel search and collect queries may be run at the time. Depending on the search location, the resulting amount of data can be quite large, putting pressure on the infrastructure. In order not to miss out on messages, the responding times of the system cannot be compromised.

5 Of course, the size of the file depends on the format. A ziped comma-separated value format will be much smaller. One week provides good

comparison of data over a number of days and also shows the different activity patterns between weekdays and weekends. Furthermore, because of the IP and infrastructural limitations, we continuously have to make way for new collections.

6 See http://www.facebook.com/privacy/explanation.php.

7 See http://www.facebook.com/policy.php.

8 See http://mattmckeon.com/facebook-privacy/.

9 Twitter's statement on privacy is available at http://twitter.com/privacy.

10 For more details, see http://pleaserobme.com/.

11 For more details, see http://urbantick.blogspot.com/2011/04/location-information-collection-creepy.html and http://ilektrojohn.github.com/creepy/.

12 Online is not equal to public. Even though data acquired through an API are branded public, it cannot be simply taken as information in the public domain. To illustrate this, we might contrast data harvested from Twitter with online resources that make data available in response to freedom of information legislation. Among others, the Guardian Data Store (http://www.guardian.co.uk/data) or data.gov.uk offers examples where governmental information is disclosed (available at http://www.facebook.com/privacy/explanation.php). More importantly, a public authority has actively decided for these data to be made available.

References

Allen, C. (1996) 'What's wrong with the "golden rule"? Conundrums of conducting ethical research in cyberspace', *The Information Society*, vol. 12, no. 2, pp. 175–188.

Auletta, K. (2009) *Googled: The End of the World as We Know It*, The Penguin Press, New York, NY.

Bakardjieva, M. & Feenberg, A. (2001) 'Involving the virtual subject', *Ethics and Information Technology*, vol. 2, no. 4, pp. 233–240.

Baker, K. & Bowker, G. (2007) 'Information ecology: open system environment for data, memories, and knowing', *Journal of Intelligent Information Systems*, vol. 29, no. 1, pp. 127–144.

Beaulieu, A. (2010) 'From co-location to co-presence: shifts in the use of ethnography for the study of knowledge', *Social Studies of Science*, vol. 40, no. 3, pp. 453–470.

Beaulieu, A. & Estalella, A. (2011) 'Rethinking research ethics for mediated settings', *Information, Communication & Society*, iFirst publication, DOI: 10.1080/1369118X.2010.535838.

Boellstorff, T. (2008) *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human*, Princeton University Press, Princeton, NJ.

Bowker, G. (2000) 'Biodiversity data-diversity', *Social Studies of Science*, vol. 30, no. 5, pp. 643–683.

Boyd, D. (2008) 'Social network sites: the role of networked publics in teenage social life', in *Youth, Identity, and Digital Media*, ed. David Buckingham, The MIT Press, Cambridge, MA, pp. 119–142.

Boyd, D. & Ellison, N. B. (2008) 'Social network sites: definition, history, and scholarship', *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230.

Carusi, A. (2008) 'Beyond anonymity: data as representation in e-research ethics', *International Journal of Internet Research Ethics*, vol. 1, no. 1, pp. 37–65.

Carusi, A. & Jirotka, M. (2009) 'From data archive to ethical labyrinth', *Qualitative Research*, vol. 9, no. 3, pp. 285–298.

Chang, O. (2011) 'Twitter hits nearly 200M accounts, 110M tweets per day', Focuses on global expansion, *Forbes Oliver Chiang*, [Online] Available at: http://blogs.forbes.com/oliverchiang/2011/01/19/twitter-hits-nearly-200m-users-110m-tweets-per-day-focuses-on-global-expansion/ (6 March 2011).

de Certeau, M. (1984) *The Practice of Everyday Life*, University of California Press, Berkeley, CA.

Dunt, I. (2009) 'Analysis: is Google Latitude a problem?', *politics.co.uk*, [Online] Available at: http://www.politics.co.uk/analysis/analysis-is-google-latitude-a-problem-s1279630.htm (20 March 2011).

Dutton, W. H. (2010) 'Reconfiguring access in research: information, expertise, and experience', in *World Wide Research: Reshaping the Sciences and Humanities*, eds W. H. Dutton & P. Jeffreys, The MIT Press, Cambridge, MA, pp. 21–39.

Dutton, W. H. & Jeffreys, P. W. (2010) *World Wide Research: Reshaping the Sciences and Humanities*, The MIT Press, Cambridge, MA.

Dutton, W. H. & Meyer, E. (2009) 'Experience with new tools and infrastructures of research: an exploratory study of distance from, and attitudes towards, e-research', *Prometheus*, vol. 27, no. 3, pp. 223–238.

Dutton, W. H. & Piper, T. (2010) 'The politics of privacy, confidentiality, and ethics: opening research methods', in *World Wide Research: Reshaping the Sciences and Humanities*, eds W. H. Dutton & P. Jeffreys, The MIT Press, Cambridge, MA, pp. 223–240.

Eagle, N., Pentland, A. & Lazer, D. (2009) 'Inferring friendship network structure by using mobile phone data', *Proceedings of the National Academy of Sciences*, vol. 106, no. 36, pp. 15274–15279.

ESRC (2008) 'Research ethics framework', [Online] Available at: http://www.esrc.ac.uk/ESRCInfoCentre/Images/ESRC_Re_Ethics_Frame_tcm6–11291.pdf (10 March 2008).

Ess, C. (2009) *Digital Media Ethics*, Polity Press, London.

Fielding, N., Lee, R. & Blank, G. (2008) *Sage Handbook of Online Research Methods*, Sage, Thousand Oaks, CA.

Frankel, M. S. & Siang, S. (1999) 'Ethical and legal aspects of human subjects research on the Internet: a report of a workshop', American Association for the Advancement of Science, Washington, [Online] Available at: http://www.aaas.org/spp/sfrl/projects/intres/report.pdf (20 February 2011).

Friedberg, A. (2004) 'Virilio's screen: the work of metaphor in the age of technological convergence', *Journal of Visual Culture*, vol. 3, no. 2, pp. 183–193.

Harrison, R. (2009) 'Excavating second life: cyber-archaeologies, heritage and virtual communities', *Journal of Material Culture*, vol. 14, no. 1, pp. 75–106.

Hine, C. (2006a) 'Databases as scientific instruments and their role in the ordering of scientific work', *Social Studies of Science*, vol. 36, pp. 269–298.

Hine, C. (2006b) *New Infrastructures for Knowledge Production: Understanding E-Science*, Idea Group Inc, Hershey, PA.

Hine, C. (2008) *Systematics as Cyberscience. Computers, Change and Continuity in Science*, The MIT Press, Cambridge, MA.

Hogan, B. (2008) 'Analyzing social networks via the Internet', in *Sage Handbook of Online Research Methods*, eds N. Fielding, R. Lee & G. Blank, Sage, Thousand Oaks, CA, pp. 141–160.

Huberman, B. A., Romero, D. M., Galuba, W. & Asur, S. (2010) *Influence and Passivity in Social Media*, HP Research Social Computing, Palo Alto, CA.

Irani, D., Webb, S., Li, K. & Pu, C. (2009) 'Large online social footprints – an emerging threat', Computational Science and Engineering, IEEE International Conference, Los Alamitos, CA, IEEE Computer Society, Vol. 3, pp. 271–276.

Jeffreys, P. (2010) 'The developing conception of e-research', in *World Wide Research: Reshaping the Sciences and Humanities*, eds W. H. Dutton & P. Jeffreys, The MIT Press, Cambridge, MA, pp. 51–66.

Jenny, B. & Raeber, S. (2008) 'Rudolf Leuzinger. relief shading', [Online] Available at: http://www.reliefshading.com/cartographers/rleuzinger.html (7 March 2011).

Kiss, J. (2009) 'Google Latitude: is the public ready for mobile tracking?', *Guardian Blog*, [Online] Available at: http://www.guardian.co.uk/media/pda/2009/feb/05/google-mobilephones (20 March 2011).

Kraak, M. J. (2010) *Cartography: Visualization of Geospatial Data*, 3rd edn, Prentice Hall, Harlow.

Krishnamurthy, B. & Wills, C. E. (2008) 'Characterizing privacy in online social networks', *Proceedings of the First Workshop on Online Social Networks*, vol. 1, pp. 37–42.

Leavitt, A. (ed.) (2009) 'The Iranian election on Twitter: the first eighteen days', *Web Ecology Project*, [Online] Available at: http://www.webecologyproject.org/2009/06/iran-election-on-twitter/ (5 March 2011).

MACOSPOL (2011) 'Mapping controversies on science for politics', [Online] Available at: http://www.macospol.eu/ (20 March 2011).

Marres, N. & Rogers, R. (2008) 'Subsuming the ground: how local realities of the Fergana Valley, the Narmada Dams and the BTC pipeline are put to use on the web', *Economy and Society*, vol. 37, no. 2, pp. 251–281.

Quantdec (2003) 'Density calculations', [Online] Available at: http://www.quantdec.com/SYSEN597/GTKAV/section9/density.htm (7 March 2011).

Rheingold, H. (2000) [1993]) *The Virtual Community: Homesteading on the Electronic Frontier*, The MIT Press, Cambridge, MA.

Robson, K. & Robson, M. (1999) 'Your place or mine? Ethics, the researcher and the internet', in *Exploring Cyber Society: Social, Political, Economic and Cultural issues* (vol. 2), eds J. Armitage & J. Roberts, School of Social, Political and Economic Science, University of Northumbria at Newcastle, UK.

Rogers, R. (2004) *Information Politics on the Web*. The MIT Press, Cambridge, MA.

Shore, J. & Warden, S. (2008) *The Art of Agile Development*, O'Reilly Media, Inc, Sebastopol, CA.

Tufte, E. (1983) *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, CT.

Viégas, F., Wattenberg, M. & Dave, K. (2004) 'Studying cooperation and conflict between authors with "history flow" visualizations', *CHI*, vol. 6, no. 1, pp. 575–582.

Weber, S. (2004) *The Success of Open Source*, Harvard University Press, Cambridge, MA.

Webmoor, T. (2008) 'From Silicon Valley to the Valley of Teotihuacan: the "Yahoo!s" of new media and digital heritage', *Visual Anthropology Review*, vol. 24, no. 2, pp. 183–200.

Woolgar, S. (ed.) (2002) 'Five rules of virtuality', in *Virtual Society? Technology, Cyberbole, Reality*, Oxford University Press, Oxford, pp. 1–23.

Wu, S., Hofman, J. M., Mason, W. A. & Watts, D. (2011) 'Who says what to whom on Twitter', *Yahoo! Research*, [Online] Available at: http://research.yahoo.com/pub/3386 (28 March 2011).

Fabian Neuhaus is a PhD candidate at the Centre for Advanced Spatial Analysis (CASA) at University College London. *Address*: Centre for Advanced Spatial Analysis, University College London, London, UK. [email: neuhaus.fabian@gmail.com]

Timothy Webmoor is a research fellow in Science and Technology Studies (STS) at the Institute for Science, Innovation and Society at the University of Oxford. *Address*: Institute for Science, Innovation and Society, Oxford University, Park End Street, Oxford OX1 1HP, UK. [email: timwebmoor@gmail.com]