International Journal of Social Research Methodology, 2013
Vol. 16, No. 3, 245–260, http://dx.doi.org/10.1080/13645579.2013.774185

Routledge
Taylor & Francis Group

Digital social research, social media and the sociological imagination: surrogacy, augmentation and re-orientation

Adam Edwards*, William Housley, Matthew Williams, Luke Sloan and Malcolm Williams

School of Social Sciences, Cardiff University, Cardiff, UK

(*Received 1 May 2012; final version received 23 January 2013*)

Technological innovation in digital communications, epitomised in the shift from the informational web (Web1.0) to the interactional web (Web2.0), provokes new opportunities and challenges for social research. Web2.0 technologies, particularly the new social media (e.g. social networking, blogging and micro-blogging) as well as the increased accessibility of the *World Wide Web* through highly portable and prevalent devices like smart phones, tablets and netbooks generates new forms of data which are of significance for social research as well as new methods and techniques for analysing this kind of data. Even though we are in the midst of this rapid innovation, it is nonetheless possible to distinguish three basic lines of argument about its current and prospective impact on social research. Some commentators suggest this innovation generates methods and data that can act as a *surrogate* for more traditional quantitative and qualitative research designs such as experiments, sample surveys and in-depth interviews. Others argue that digital communication technologies *re-orientate* social research around new objects, populations and techniques of analysis. It can also be argued that digital social research *augments*, but needs to be used in conjunction with, more traditional methods. C. WrightMills' classic statement of *The Sociological Imagination* is used to clarify the distinctive contribution of digital social research; what can it do that traditional methods cannot in understanding how social relations are constituted, how they can change and how they generate social identities. It is argued that digital social research, particularly the analysis of new social media, is distinctive in capturing naturally occurring or 'user-generated' data at the level of populations in real or near-real-time. Consequently, it offers the hitherto unrealisable possibility of studying social processes as they unfold at the level of populations as contrasted with their official construction through the use of 'terrestrial' research instruments and curated data-sets. Realising this research potential entails the development of digital 'observatories' such as the *Cardiff Online Social Media ObServatory*. The paper concludes with a discussion of the political and ethical, as well as the technological, implications of observatories, focusing in particular on tensions between the 'panoptic' and 'synoptic' powers of digital observatories and the allied possibilities of a 'signature science'.

Keywords: empirical crisis; social media data; surrogacy; augmentation; re-orientation

*Corresponding author. Email: EdwardsA2@Cardiff.ac.uk

Introduction

> One key assumption here is that understanding the imbrications between digitization and politico-economic process requires recognizing the embeddedness of digital space and resisting purely technological readings of the technical capacities entailed by digitization. One of the subjects... to address is the as yet underdeveloped analytics for understanding digital technology from a social science perspective. (Sassen, 2006, p. 329)

In their account of the 'coming crisis of empirical sociology', Savage and Burrows (2007) argue that, in the previous four decades, social scientists were able to claim a distinctive expertise in investigating social relations through such methodological innovations as the sample survey and the in-depth interview. Since the advent of digital communications technologies, this claim has been compromised by the proliferation of 'social transactional data' generated, owned and increasingly analysed by large commercial organisations as well as government departments. The advent of 'big and broad data' on, for example, retail transactions, telephone communications, financial expenditure and insurance claims in addition to the digital production of official, 'curated', data-sets such as the census of populations, general household surveys, police recorded crime, victim of crime surveys and labour market surveys, provokes an existential question for academic sociology; is it, 'becoming less of an 'obligatory point of passage' for vast swathes of powerful agents ... if so, how can the discipline best respond to this challenge?' (Savage & Burrows, 2007, p. 886).

Concern about the marginality of academic sociology is arguably compounded by the subsequent explosion of new social media communications, such as social networking sites, the 'blogosphere' and the increasing popularity of micro-blogging or, named after the most renowned micro-blogging service, 'tweeting'. Such Web2.0 technologies enable direct communication between social elites, from electoral candidates and entertainers to Olympians and other sports stars and 'digital publics', whilst also registering mass sentiments about political campaigns, performances and sporting events. Significantly, these technologies also facilitate the mass communication of 'user-generated content', given the portability and increasing prevalence of smart phones, net-books and tablets that facilitate rapid access to social media on the move. Such technologies generate a form of mass self-report data about social media users' perceptions of particular events. Social and computational researchers have already begun to mine and 'repurpose' this naturally occurring socially relevant data in their 'predictive' efforts. Tumasjan, Sprenger, Sandner, and Welpe (2010) were able to measure twitter sentiment in relation to candidates in the German general election concluding that this source of data was as accurate at predicting voting patterns as traditional polls. Again, mining the twittersphere, Asur and Huberman (2010) were successful at correlating the sentiment expressed about movies with their revenue, claiming that this method of prediction was more accurate than the gold standard Hollywood Stock Market. Beyond social networks, Ginsberg et al. (2009) successfully correlated flu-based search terms entered into the Google engine with visits to the local doctor to epidemiologically trace the spread of the disease across the USA. Another notable example is the wealth of social media communications about major incidents of civil unrest, such as the riots in English cities during August 2011 (Procter, Viz, & Voss, 2013). These studies illustrate the potential significance of social media technologies for facilitating the

generation and analysis of 'naturally occurring' mediated data as contrasted with findings from experiments, surveys and in-depth interviews which are necessarily the artefacts of social researchers (Cicourel, 1964).

However, of the studies mentioned above, few have been replicated with equally positive results calling into question sweeping statements about the birth of an entirely new way of social data generation and collection. Furthermore, research into the riots of August 2011, particularly the investigation led by the Guardian and the London School of Economics and Political Science, *Reading the Riots* (Guardian 2011, Procter et al., 2013), has identified some of the methodological limitations of social media communications as data for social research. This study demonstrated the low fidelity of social media communications as representations of popular sentiment about the riots and their causes. It identified the prevalence of misinformation, pranks, rumour and sarcasm. Further, unlike the front-sheet of a sample survey, it is difficult to ascertain with any precision the demographic characteristics of micro-bloggers as contrasted with survey respondents or qualitative interviewees; there is little in the 140 characters of a tweet formulation, about the age, gender, ethnicity, social class and other attributes of the tweeter in question. While it may be possible to use proxies for some demographic attributes, as part of a 'signature science' to make inferences from social media (see 're-orientation', below and Housley, Edwards, Williams, & Burnap, forthcoming), it is questionable as to whether the analysis of this data stream can act as a *surrogate* for traditional social research methods. It remains unclear how social media as 'data' could, for example, compete with the much greater fidelity of surveys of victims of crime or self-report studies of offending behaviour or epidemiological surveys of public health trends.

At the very least, initial exercises in the use of social media as sources of data for social science give pause for thought about the crisis of empirical sociology and the continued significance of traditional, 'terrestrial', research methods; the demise of the survey and the interview are much exaggerated. The continuing centrality of traditional methods is brought into greater focus still, when the significance of 'context' and contextualised insight is considered. Surely, one of the principal claims to distinctive expertise that academic sociology can make is its ability to place social relations in their specific historical and cultural contexts, as epitomised in ethnographic methods of participant observation. Of course, Web2.0 technologies have given rise to online social relations including the avatars of participant observers in Virtual Worlds such as 'second life' (Williams, 2006). The point is that it is not at all clear that data derived from social media communications provide social researchers with an alternative to ethnographic immersion in offline contexts as a means of understanding these contexts, much less of broaching the greater challenge of investigating the interaction between offline and online social relations.

Rather than regarding social media data analysis as a surrogate for traditional methods of inquiry, therefore, the experience of digital social research to date suggests the need for an account of how it might augment, and be augmented by, traditional social research methods. In addition to considering the scope for such augmentation, it is argued here that what is really distinctive about social media data is the production of naturally occurring or 'user-generated' data at the level of populations in real or near-real-time. Consequently, it offers the hitherto unrealisable possibility of studying social processes as they unfold at the level of populations as contrasted with their retrospective official and academic construction through the use of terrestrial research methods. Given the low fidelity of such data, however, we are,

as yet, uncertain about how the populations so represented through social media communications can be understood or analysed using conventional techniques of quantitative data analysis. In addition, the sheer volume of social media communications places significant logistical limitations on the use of qualitative data analysis techniques. On the other hand, it may be that social media communications provoke a re-orientation of social science around new kinds of objects, different kinds of populations and alternative techniques of analysis. In support of this argument it is helpful to conceptualise the significance of the new social media for social science in terms of the kinds of research strategy, design and data that it facilitates.

Distinguishing the contribution of the new social media to social science

To clarify the distinctiveness of social media and its analysis from more established social research, it is helpful to reprise the programme underpinning C. Wright Mills' concept of the 'sociological imagination'. It will be recalled that Mills posed three basic types of question driving this programme:

(1) What is the structure of this particular society as a whole? What are its essential components … how does it differ from other varieties of social order …
(2) Where does this society stand in human history? What are the mechanics by which it is changing? …
(3) What varieties of men and women now prevail in this society and in this period and what varieties are coming to prevail? (Mills, 1959, p. 13).

In these terms, the computational analysis of new social media in real and near-real-time[1] enables new ways of approaching the empirical analysis of social organisation, social change and the integration of the individual into collective life. For example, digital network analysis of social media communication is key to identifying new conduits for influence, the emergence of digital elites (e.g. the 'Twitterati') and hitherto occluded identifiable measures of network capital. Replicable analyses of social media communications and the behaviour of digital elites and publics in relation to key events provide empirical opportunities and access to big and broad data streams. Finally, the ways in which we empirically document prevailing forms of 'human types' are transformed through the agential modelling of online behaviour within digital societies. Of course this new empirical programme for digital sociology has to be contextualised in terms of theory building and its interrelationship with methodological innovation and empirical data collection and analysis (Housley et al., 2013). In the following sections of the paper we will discuss how

Table 1. The distinctiveness of new social media analysis in relation to more traditional research strategies, design and data.

		Research design/data	
		Locomotive	Punctiform
Research Strategy	Intensive	e.g. Ethnography/Participant Observation	e.g. Qualitative interviewing
	Extensive	Social media analysis (capturing naturally occurring data in real time at the level of populations)	e.g. Surveys and experiments

each of the classic questions relates to three methodological positions in relation to social media data, analysis and representation. However, before doing so we need to identify these three positions.

Table 1 provides a diagrammatic representation of this distinctive contribution to social research methodology. Methodologists often distinguish intensive research strategies that capture the locomotion of social relations 'in process', and extensive strategies that capture the structure of social relations at particular moments and are therefore 'punctiform' in providing a snapshot of these relations (Sayer, 1992, pp. 156–157, 241–251). In turn, traditional social research is characterised in terms of a trade-off between, on the one hand, extensive and punctiform research which captures variation at the level of populations but only at specific moments and only in retrospect (e.g. national surveys), and on the other, intensive research strategies that capture social processes but only in very specific social contexts or amongst particular social groups (e.g. workplace ethnography). In these terms, the 'game-changing' qualities of new social media as both a source of data for, and an instrument of, social research is that it offers the promise of extensive research into social processes.

The methodological matrix presented in Table 1 affords an opportunity to frame our understanding of new social media analysis in terms of extensive/intensive strategies, locomotive/punctiform design and other more traditional strategies of data capture and interpretation. In these terms, social media analysis marks a significant departure from social research methods to capture naturally occurring data at the level of populations[2] in near and real-time. This presents social scientists with some major opportunities given that social media communications can capture and facilitate the analysis, including computation, of the sentiments (Thelwall, Buckley, & Paltoglou, 2011), tensions (Williams et al., 2013) and topics (Beer 2012) that are being uploaded by users unmediated by the constructions of official and academic social research. Again, the population level characteristics of social media data afford social scientists opportunities, namely accessing big and broad data streams that can be linked to key events over time, thereby rendering populations visible and thinkable in both their locomotive (in motion) and reactive states. For example, the ways in which different populations convey content relates to major incidents of civil unrest, foreign policy interventions, global sporting occasions and disasters and states of emergency. The analysis of social media also promises access to hitherto difficult to reach, if not invisible, populations. It is well established, for example, that young male residents of urban neighbourhoods are systematically underrepresented in conventional survey methods; the 'boyz in the 'hood' make for poor respondents to general household surveys. The advent of smartphones on cheaper pay-as-you-go services resulted in over 97% of the public owning mobile phones in 2009, with just under half of these using smartphone functions in 2011 (Dutton & Blank, 2011). This unpredicted access has seen the socio-economic digital divide close rapidly, being filled by excluded and disenfranchised youth. Finally, a major promise of social media analysis, underpinned by computational methods, is the advent of real or near-real-time analysis; the world in motion. This represents the opportunity to reframe the ways in which sociology is conducted in terms of revisiting anticipation, modelling and prediction for the purposes of more empirically robust, ethically and politically defensible early interventions against major social problems e.g. the social impact of environmental hazards and catastrophes such as earthquakes (Sakaki, Okazaki, & Matsuo, 2010) and extreme weather

(Goodman, 2012), health epidemics (Ginsberg et al., 2009) and the diffusion of inflammatory racist content (Williams et al., 2013).

These claims for the unique promise of social media analysis provoke, in turn, a discussion of what it can and cannot realistically deliver given the quality of social media communications as social research data when contrasted with the findings that can be generated through the use of more conventional methods of inquiry. One means of reflecting on what it is that social media analysis can and cannot deliver for a sociological imagination is to pose a series of supplementary questions:

(1) Can social media analysis act as a surrogate for conventional research methods, one that better accomplishes the aims and objectives of these conventional methods in investigating social structure, change and identity?
(2) Is social media analysis better understood as a means of augmenting, rather than supplanting, conventional research methods for investigating social structure, change and identity?
(3) Does social media analysis provoke a re-orientation of social science around new objects, populations, techniques and concepts for investigating social structure, change and identity?

Surrogacy

To clarify the significance of social media as data for social research, it is helpful to question what insights it can provide that other sources of data cannot. Does it help us to investigate conventional research questions, for example about social stratification, in ways that are superior, more efficacious, to those afforded by conventional research methods, such as sample surveying. Is it the case that the 'big and broad datasets' generated by social media empirically document social relations and populations hitherto inaccessible by surveying, qualitative interviewing and participant observation and in ways that are more 'insightful' and 'reflective' of the contemporary social milieu. To elaborate this point further and relate it to Mills' concept of the sociological imagination, we consider the efficacy of social media analysis for investigating questions of social structure and organisation, social change and social identity.

Surrogacy: social structure and organisation

Notwithstanding Savage and Burrows' (2007) concerns over the perishable qualities of conventional research methods in a world awash with transactional data, there are several reasons for caution in writing the obituary of the survey and the interview as the principal means of realising Mills' empirical programme for social research. As noted above, social media communications yield data that is of low fidelity when contrasted with the demographic details that survey 'cover sheets' provide or the biographic insights that qualitative interviewing can generate. Further, unlike household survey designs, such as the Census of the population, or other official data-sets such as police recorded crime, there is limited insight into the geographic location of new social media users, not least as the default setting of the GPS tracking device in smart phones is switched to 'off' and only in a fraction of cases (less than 1%) do users switch on this locater. Furthermore, most social media sites impose limits on harvesting from their application programme interfaces

(APIs). The amount that can be harvested varies by provider. Twitter for example, have several services including the Spritzer (a randomly sampled 1% of all tweets), the Gardenhose (a randomly sampled 10%) and the Firehose (all public tweets). However, the provision of streams above one percent is rare. Most free and commercial software programs available that facilitate API access are subject to 'rate limitations' making missing data a potential problem. This often means, in the case of Twitter at least, that researchers will only get a very small fraction of the one % of tweets that are geo-located. So, in summary, it is not abundantly clear who social media users are, particularly those using micro-blogging sites, where they are from or where they are located whilst using social media. As such many of the core demographic questions driving sociological research into patterns of social stratification, inequalities and division around issues of age, gender, ethnicity and social class cannot be posed using social media in the same way they can be through more conventional methods.[3] To take crime as an issue of acute interest to both the general public and to the social research community, analysis of social media would make a poor surrogate for the insights into the relationship of victimisation and social inequality that is possible through linking household survey data-sets, such as the British Crime Survey and Census (Hope, 1997; Trickett, Ellingworth, Hope, & Pease, 1995).

Surrogacy: social change

From a socio-historical perspective, new social media data may provide invaluable insight into societal change if it can be classified, stored and analysed systematically in ways that are recoverable for future social scientific inquiry. Big and broad data recovered from publically available social media data in relation to key events over time (e.g. the Olympics, wartime, civil unrest or elections) offer unprecedented insight into the responses of populations to the unfolding logos of world and societal processes. Furthermore, population and group reactions to events may also include responses to and the use of new technologies. Yet alone these data cannot act as a surrogate for conventional methods of investigating social change. Whilst there is a clear affinity between 'blogs' and the personal diaries that have proven so critical for historical research, social media communications tend to be more parsimonious in content, as epitomised in the 140 character formulation of 'tweets'. Of course these communications can provide an archive of prevalent sentiments, opinions and topics of interest, much of which are routinely captured and displayed by micro-blogging sites that analyse what is 'trending' at particular moments. This is also aided by the convention of placing a hashtag ('#') in front of particular sentiments, opinions and topics which enables their cross-referencing and propagation or 're-tweeting'. In relation to this, there is an emerging social science of online trends in the use of social media communications, for example to propagate, corroborate or contest 'rumours' (Procter, Crum, Karstedt, Voss, & Cantijoch, 2013). Even so and notwithstanding the significant demands on high-powered computation implied by the capture, analysis and archiving of social media communications (see 'augmentation', below), they provide a poor surrogate for conventional methods of investigating processes of social change, certainly in the offline world.

Surrogacy: social identity

Given the low fidelity of social media communications for indicating the characteristics and qualities of offline populations, they are also a poor surrogate for conventional methods of investigating the 'varieties of men and women' inhabiting particular societies. The question for surrogacy in relation to the study of identity formation rests upon the idea that one can use social media to identify data that relate to conventional categories of social identity in ways that are equivalent to terrestrial methods which thereby render these conventional methods obsolete. The promise of social media analysis is that it can provide extensive and locomotive profiles of populations. But again, due to issues of fidelity, this is not always the case. Furthermore, identity formation and the promotion of specific 'human types' within social formations has benefited from ethnographic and other forms of traditional locomotive and intensive inquiry (Hammersley & Atkinson, 1983). However, through the use of signature proxies for age, class, gender and ethnicity social media analysis may well be fruitfully used in order to augment traditional strategies of data collection, analysis and representation.

Augmentation

If, therefore, social media analysis does not provide a superior means of investigating conventional research questions, it may nonetheless be able to augment conventional methods in posing these conventional questions; to improve their accessibility and insight into hitherto under-represented social groups. For example, if the 'boyz in the "hood"' make for poor household survey respondents, might they be better accessed through social media and, if so, how might this data relate to findings from survey research, ethnography and other more conventional methods?

Augmentation: social structure and organisation

A preoccupation of conventional research into social structure and organisation is the impact of division and inequality on problems of public health, education, imprisonment and violence etc. exemplified recently by the 'spirit level thesis' that societies characterised by gross inequalities of household income also exhibit lower levels of life expectancy and educational attainment with elevated levels of violence, imprisonment and mental ill-health (Wilkinson & Pickett, 2009). If the low fidelity of social media data cannot provide a surrogate test of such epidemiological research, it may nonetheless augment it, for example by indicating the prevalence of attitudes and sentiments that validate or challenge such inequalities.

A stronger case for the augmentation of conventional research by social media analyses is in relation to investigation of the purported 'networked society' (Castells, 1996), specifically the proposition that social structures are becoming less hierarchical, not least as a consequence of the revolution in digital communications and the allied mobility of populations across social groups and territories. Social media analyses can complement conventional methods of investigating the density of offline social networks, the identification of influential actors and organisations or 'nodes' within these networks and comparative differences in the structure of elite and more open-ended networks (Scott, 1991) through examination of online

networks of communication; do, for example, online networks exhibit similar patterns of openness and closure, of the propagation or disruption of rumours or received wisdom, in relation to interests shared by offline networks (e.g. political and economic controversies, electoral preferences and consumer tastes etc.).

Yet, as mentioned above in relation to arguments about surrogacy, social researchers still require terrestrial methods and access to curated data-sets in order to investigate any putative relationship between online communications and offline behaviour (e.g. social media discussion of political protests and elections, industrial disputes and sporting events that anticipate or incite conflict and the actual realisation of this conflict during demonstrations, on picket lines and at sporting contests). Further thought needs to be given to how the characteristics of online communication amongst actors using social media can be cross-referenced with offline data on the class, ethnicity, gender, age and location etc. of social actors.

Augmentation: social change

Although social media data can be criticised for its uncertain relationship to conventional social categories and cannot provide a surrogate analysis of the social trends identified through surveys or ethnographic studies premised on these categories, where the categorisation of the populations and groups in question is more evident, it may still augment the study of social change through use of 'proxy' data. In addition to the use of digital databases of names, as an imperfect proxy for gender A-Z street names might be used as another, imperfect, proxy for location. For example, epidemiological research into the impact of income inequalities on social problems over time and from one age cohort to another could be augmented through linguistic and metric analyses of social media communications indicating the content and frequency of competing sentiments and attitudes about inequalities, particularly if these corresponded to particular places and issues identified in conventional research into offline behaviour. Obviously this kind of analysis cannot refer to historical data pre-dating the onset of Web2.0 communications but the archiving of social media data could enable investigations of any co-evolving relationships between trends in offline social relations and their signature in online communications (e.g. through the identification of prevalent, 'trending', sentiments and hashtags). Analysis of the networked structure of social media communications can also have a temporal dimension as measures of density and influence in online social networks are more amenable to continual measurement, to reveal how changeable or 'locomotive' they are, relative to the more retrospective, 'punctiform', analyses of offline networks. An important implication of this is that social media analysis can augment conventional research into these networks; are online social networks more permeable; less insular and open to restructuring than offline networks which share the same interests? An obvious example here, of major significance for representative democracy, is the openness of online networks used for the mass communication of electoral campaigns as contrasted with the relative insularity of the print and broadcast media; can social media have a catalytic effect on political change in liberal democracies and, for that matter, in polities where there is greater restriction on conventional mass media?

Augmentation: social identity

As noted above, names and streets may be used as proxies for the higher fidelity demographic data routinely collected through surveys or observed through ethnographies and other qualitative methods. This is particularly significant in relation to one of the key prospective advantages of treating social media data as relevant for conventional social research questions; the prospect of accessing hard to reach and hard to hear groups that are routinely under-represented in household surveys or only captured by qualitative methods in highly specific contexts and moments. In order to realise this potential, however, it is necessary to shift the empirical focus of social media analysis from meta-data on social media users ('the cover sheet') toward the substantive content of social media communications (e.g. the 140 word formulation of a 'tweet'). In this regard, social media analysis can both augment and be developed through conventional linguistic and conversation analysis where membership categories (e.g. named individuals, social groups, locations) are related to membership categorisation devices (particular events or topics) and predicates (including extreme case formulations, 'totally', 'completely' etc. and degradation terms) to interpret the cultural meaning of communication in context (Sacks, 1992; Housley & Fitzgerald, 2002, 2009). This approach has been used to indicate the interplay between racist identities and social media communications about events in professional football in England (Williams et al., 2013). In turn, the indication of identity through these kind of methods can augment conventional research into identity formation for example in research on tension indication in the policing of neighbourhoods and major public events which draws upon official data-sets, such as police recorded crime statistics and findings from victimisation surveys along with the qualitative intelligence from patrol officers and other key informants (Williams et al., 2013).

Re-orientation and signature science

Re-orientation can be understood to begin at the end point of augmentation discussed above. The argument here is that social media communications raise new social science research questions that are different from more conventional questions. Furthermore, different forms of analysis and concepts are required to answer these questions. In order to take best advantage of the new affordances of social media communications, social and computational researchers are building platforms that are able to harvest, archive and analyse these data. The Cardiff Online Social Media Observatory ('COSMOS', see Burnap et al; this issue) is such a platform offering a digital data observatory which takes advantage of the concept of 'signature science' – the measurement of key variables of interest via digital proxies.

Re-orientation: structure, organisation, change and identity

Data derived from social media communications invite social researchers to think differently about questions in relation to social structure and organisation, change and identity. For instance, the non-pyramidal and non-hierarchical structures inherent in online communications (Spears & Lea, 1992) question the relevance of more conventional individual attributes (race, age, class, gender etc.) that have taken prime place in terrestrial social research. Therefore questions about social structure,

change and identity are re-orientated to the new ways in which digital agents organise, change and identify. The work of Knorr-Cetina (2001) on 'objectual practice' becomes relevant when examining the structure, change and identity of new 'digital publics'. The concept of object orientated sociality, where actors convene around particular knowledge or cultural objects, such as parenting, job seeking, dating, following a sports team, celebrity or entertainer etc. may be more useful in understanding new forms of social organisation, change and identity than more traditional notions of gender, class and race. In relation to social change, 'traditional' notions of equality may be less relevant to new 'digital publics' where alternative forms of facilitators and inhibitors to mobility may be in operation; characteristics of the digital online elite (e.g. the 'twitterati') may not map onto the characteristics of the traditional terrestrial elite. Knowledge, access and connections to online objects may be more important in facilitating mobility than an actor's gender, class or race. Similarly, how members of the 'digital public' identify or are identified may relate more to the objects they connect with than their terrestrial attributes. Furthermore, the nature of these connections (e.g. frequency and sentiment of communications) become the focus of a 'signature science' which begins to identify actors and the ways they organise and change via 'behaviour-metrics' – the behavioural characteristics and patterns of actors in online environments: information facilitators/inhibitors, bridges between online communities, opinion formers etc. The absence of traditional categories and attributes of individuals is therefore less of a concern for a signature science which aims to derive variables of interest from empirical observations of behaviours in either their manifest or latent forms.

While these theoretical concepts may not be 'new' per se, the re-orientation of techniques and methods to study them via social media are. Our practices as social researchers are stretched as we attempt to grapple with socially relevant data that are naturally occurring in real time at the level of populations. This requires changes in practice to adapt existing methods and adopt new ones from disciplines that at first glance may not seem cognate. Computer science provides tools and architectures that ameliorate some of social sciences' contemporary problems with big and broad data-sets. Cooperative working provides the space within which to explore the computational codification of social science concepts rendering vast amounts of data analysable and visualisable in conventional and unconventional ways. Collaborative algorithm design involves the combination of measurement; construct validation; and interpretation through an iterative and interactive process. For example, operationalising social science theoretical propositions through the practical design and codification of computational tools and methods, which in turn produce results that are subjected to further critical interpretation and refinement.

It is at the interface between disciplines that digital data observatories such as COSMOS become realisable, and that data from social media networks come to be readily analysable and relevant to the problems with which social scientists concern themselves. Such observatories also afford opportunities for the scientific replication of research via their digital infrastructure. Whilst not a re-orienting manoeuvre for quantitative social research, this provision is novel to qualitative social research. The key inhibitors to re-analysing and re-purposing qualitative data are well rehearsed and include questions of differences in epistemological approach between the data generator and re-user/re-purposer, issues of ethics including the continuation of informed consent from the original work to its secondary analysis and representation, and issues of reflexivity and the production of sufficient

meta-data to facilitate a 'credible', 'transferable', 'dependable' and 'confirmable' secondary account (Dicks, Mason, Williams, & Coffey, 2006; Lincoln & Guba, 1985). Digital research observatories like COSMOS and platforms like *My Experiment* offer tools and research workflow models that nullify some of these inhibitors enabling the use of mixed methods in studies aimed at the replication of social research findings. For example, through various annotation tools, digital observatories allow meta-data about how the research in question was conducted and analysed to be routinely and systematically recorded in ways that assist in the re-production of research designs and the comparison of subsequent results.

In anticipation of ethical concerns over the potential uses of digital observatories for the purposes of intrusive surveillance, it must be acknowledged that the capacity to rapidly harvest and link or 'mash' social media communications (which may be anonymous) with other digital data sources which subsequently reveal the identity of users, is a significant danger, particularly in more authoritarian contexts. How such 'panoptic' power can be policed or inhibited is a major issue in the quickly evolving world of the interactional web. Even so, concern over panoptic power ought not to crowd out an appreciation of the re-orientation of social research in support of the 'synoptic' powers of the interactional web. The 'viewer society' (Mathiesen, 1997), in which the many are able to view the few (whether these be political and economic elites or front-line public order police officers) promises hitherto unimaginable opportunities for challenging elite constructions of social problems and opening out opportunities for a more democratic and deliberative approach.

Conclusion: collaborative algorithm design and signature science

During the course of this paper, we have considered the potential for social media analysis to substitute, augment and re-orient traditional social research methods. We made arguments that questioned the notion that social media communications and the data derived from them constitute a surrogate to more conventional survey and interview methods. The fidelity of the data is too low and it is currently difficult to extract attribute data that social scientists use on a daily basis (age, gender, race, class, etc.). We instead argued that this new source of naturally occurring socially relevant data can augment existing methodological practices, by affording insights into the behaviours and perceptions of hard-to-hear groups that routinely avoid national surveys and the like, and by providing an ability to monitor digital populations in a locomotive fashion. In the final part of the paper, we claimed that social media also has the potential to re-orientate the social scientist and their practices. New ways of organising, mobilising and identifying are emerging within online environments which question the relevance of traditional terrestrial characteristics and attributes ascribed upon individuals. Furthermore, signature science makes observable online behaviours which may empirically prove more meaningful in explaining social life in a networked world, while digital observatories afford new opportunities for making sense of this data and future replication.

However, the full realisation of a signature science associated with digital observatories is reliant upon the development of social media itself, in particular the retention of open access APIs, the software that enables digitally stored data to be fed into other software packages (e.g. the APIs that feed micro-blogs, 'tweets', into packages that analyse their linguistic content, measure their frequency and reveal

the structure of online communications, such as the volume of micro-bloggers involved in a communication and the role of particularly influential bloggers in propagating or disrupting this communication, see Proctor et al., 2013). Whilst Twitter has provided publically accessible data streams there is no guarantee that this will always be the case as social networking sites find new ways to secure income streams and exploit the commercial properties of this data. To this extent, the political economy of social media and digital data more broadly is a key concern and social scientists need to explore the provision of stable digital data feeds if they are to take full advantage of big and broad data.

A further concern is the extent to which the analysis of social media relies on 'off the shelf' devices and tools developed by marketing companies whose algorithmic aspects remain 'black boxed', opaque and hidden from inspection. In terms of replicable social science, the repetition of analyses through work flow platforms (such as COSMOS and My Experiment) on comparable data-sets needs to be aligned with the inspection of the algorithms that drive different digital tools and devices (as in the case of different sentiment analysis packages). Are we to expect social scientists to lobby commercial developers to render visible their algorithms and so, in effect, to hand over commercial secrets for scientific purposes? In some cases quite possibly; but we would also suggest that sociologists need to engage with a process of collaborative algorithm design in ways that draw upon the well-established field of computer supported co-operative work (Button & Sharrock, 1996). Furthermore, such collaborative work within the domain of 'open science' critically depends on open societies where matters relating to data use, surveillance, panoptic and synoptic power are coming to the fore as ethical and political concerns. To this extent digital observatories represent disruptive technologies that can open up public online data to inspection and analysis in ways that are transparent and potentially rigorous; such processes need to be underpinned by ethical concerns in relation to storage, use and anonymity (Williams et al., 2007). To this extent the design and establishment of observatories represents a mutual intertwining of method, engineering and ethics in an age of digital data.

Acknowledgement

This work is based on the project 'Digital Social Research Tools, Tension Indicators and Safer Communities: a Demonstration of the Cardiff Online Social Media ObServatory (COSMOS)', which was funded by the UK Economic and Social Research Council under the Digital Social Research Demonstrator Programme (Grant Reference: ES/J009903/1).

Notes

1. Near-real-time analysis better reflects the practical limitations imposed by computational power when processing big and broad data streams.
2. That is of course if we are considering social media users as the population of interest. See Housley et al. (forthcoming) for an overview of the limitations of social media data for quantitative analysis.
3. Although the use of digital proxies for gender can be ascertained utilising 'name type' databases.

Notes on contributors

Adam Edwards is a senior lecturer and director of the Centre for Crime, Law and Justice at the Cardiff School of Social Sciences, Cardiff University. His current research interests include the emergence of 'smart cities' and their impact on urban governance. He recently

edited the special issue of the European Journal of Criminology on 'Urban Security in Europe' (with G. Hughes, 2013). He is the co-editor of Crime control and community (with G. Hughes, Willan, 2002) and Transnational organised crime (with P. Gill, Routledge, 2006) and co-author of Digital society: Theory, method and data (with W. Housley, M. Williams and P. Burnap, Sage, 2015). He is currently co-investigator on the 'Supporting Empirical Digital Social Research for the Social Sciences with a Virtual Research Environment' project funded by JISC and the 'Digital Social Research Tools, Tension Indicators and Safer Communities: a demonstration of the Cardiff Online Social Media Observatory (COSMOS)' ESRC DSR demonstrator project.

William Housley, PhD, DSc Econ. is currently reader in Sociology at the Cardiff School of Social Sciences, Cardiff University. He has published widely in the areas of social research methods, sociology of language and interaction, ethnomethodology, interactionism and collaborative teamwork. He is the author of Interaction in multidisciplinary teams (Ashgate, 2003); co-author of Interactionism (with Paul Atkinson, part of the BSA Millennial Series, Sage, 2003), Contours of culture: Complex ethnography and the ethnography of complexity (with Paul Atkinson and Sara Delamont, Alta Mira Press, 2008), Advances in membership categorization analysis (with Richard Fitzgerald, Sage, 2014) and Digital society: Theory, method and data (with A. Edwards, M. Williams and P. Burnap, Sage, 2015). He is currently co-investigator on the 'Supporting Empirical Digital Social Research for the Social Sciences with a Virtual Research Environment' project funded by JISC and the 'Digital Social Research Tools, Tension Indicators and Safer Communities: a demonstration of the Cardiff Online Social Media Observatory (COSMOS)' ESRC DSR demonstrator project.

Matthew Williams is a senior lecturer in Criminology at Cardiff University. He is the author of Virtually criminal published by Routledge which was shortlisted for the Philip Abrams Memorial Prize of the British Sociological Association. He recently edited the special issue of Policing and society: Policing cybercrime: Networked and social media technologies and the challenges for policing. With William Housley, Adam Edwards and Pete Brunap he is writing Digital society: Theory, method and data. He is currently co-investigator on the 'Supporting Empirical Digital Social Research for the Social Sciences with a Virtual Research Environment' project funded by JISC and principal investigator for the 'Digital Social Research Tools, Tension Indicators and Safer Communities: a demonstration of the Cardiff Online Social Media Observatory (COSMOS)' ESRC DSR demonstrator project.

Luke Sloan is a lecturer in quantitative methods at the Cardiff School of Social Sciences. His research interests include: Exploring potential linkage between naturally occurring and curated data; adapting traditional social science methodology for use on social media; and student understanding and learning of quantitative methods in the social sciences. His recent publications include contributions to the British Journal of Social Work (doi: 10.1093/bjsw/bcs190), Children and Youth Services Review (34/8), Social and Public Policy Review (6/2) and Party Politics (doi: 10.1177/1354068811436045).

Malcolm Williams is the director of the School of Social Sciences at Cardiff. His research interests are principally in methodological issues in the social sciences, particularly causation, probability and objectivity. He recently published a jointly authored book (with Gayle Letherby and John Scott, Objectivity and subjectivity in social research (Sage). He is also a co-investigator on the 'Digital Social Research Tools, Tension Indicators and Safer Communities: a demonstration of the Cardiff Online Social Media Observatory (COSMOS)' ESRC DSR demonstrator project.

References

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media, CoRR*. Retrieved July 16, 2012, from http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf

Beer, D. (2012). Using social media data aggregators to do social research. *Sociological Research Online, 17*, 10. http://www.socresonline.org.uk/17/3/10.html 10.5153/sro.2618

Button, G., & Sharrock, W. (1996). Project work: The organisation of collaborative design and development in software engineering. *Computer Supported Cooperative Work (CSCW), 5*, 369–386.

Castells, M. (1996). *The rise of the network society: Information age: Economy, society, and culture volume. 1* (Information Age Series). London: Blackwell.

Cicourel, A. (1964). *Method and measurement in sociology.* New York, NY: The Free Press.

Dicks, B., Mason, B., Williams, M., & Coffey, A. (2006). Ethnography and data re-use: Issues of context and hypertext. *Methodological Innovations Online, 1*, 33–46.

Dutton, W. H., & Blank, G. (2011). *Next generation users: The internet in Britain, oxford internet survey*, available at: SSRN: http://dx.doi.org/10.2139/ssrn 1960655

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*, 1012–1014.

Goodman, J. (2012). Crime and socioeconomic conditions: Evidence for non-cultural domain specificity in evolutionary forensic psychology. *Aggression and violent behaviour, 17*, 523–526.

Guardian. (2011). Reading the riots: Investigating England's summer of disorder. Retrieved August 31, 2012 from http://www.guardian.co.uk/uk/series/reading-the-riots.

Hammersley, M., & Atkinson, P. (1983). *Ethnography: Principles and practice.* London: Tavistock.

Hope, T. (1997). Inequality and the future of community crime prevention. In S. P. Lab (Ed.), *Crime Prevention at a Crossroads* (pp. 143–158). American Academy of Criminal Justice Sciences Monograph Series. Cincinnati, OH: Anderson.

Housley, W., Edwards, E., Williams, M., & Burnap, P. (forthcoming). *Digital society: Theory, method and data.* London: Sage

Housley, W., & Fitzgerald, R. (2002). The reconsidered model of membership categorization analysis. *Qualitative Research, 2*, 59–83.

Housley, W., & Fitzgerald, R. (2009). Membership categorization, culture and norms in action. *Discourse & Society, 20*, 345–362.

Knorr-Cetina, K. (2001). Objectual practice. In T. R. Shatzki, E. Von Savigny, & K. Knorr Cetina (Eds.), *The practice turn in contemporary theory.* Abingdon: Routledge.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Newbury Park, CA: Sage.

Mathiesen, T. (1997). The viewer society: Michel foucault's "panopticon" revisited. *Theoretical criminology, 1*, 215–234.

Mills, C. W. (1959). *The sociological imagination.* Oxford: Oxford University Press.

Procter, R., Crum, J., Karstedt, S., Voss, A., & Cantijoch, M. (2013). Reading the riots: What were the police doing on Twitter? *Policing and Society.* Advance Online Access.

Procter, R., Viz, F., & Voss, A. (2013). 'Reading the riots on twitter: Methodological innovation for the analysis of big data'. In W. Housley, A. Edwards, M. L. Williams, & M. Williams (Eds.), *Computational social science: Research, design and methods. International Journal of Social Research Methods, special issue, 16.*

Sacks, H. (1992). *Lectures in conversation, volumes I and II.* Oxford: Blackwell.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors, In *Proceedings of the 19th international conference on world wide web.* New York, NY: ACM Press.

Tumasjan, A., Sprenger, T., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proc. of 4th ICWSM*, 178–185. AAAI Press.

Sassen, S. (2006). *Territory, authority, rights: From medieval to global assemblages.* Princeton, NJ: Princeton University Press.

Savage, M., & Burrows, R. (2007). The coming crisis in empirical sociology. *Sociology, 41*, 885–899.

Sayer, A. (1992). *Method in social science: A realist approach* (2nd ed.). London: Routledge.

Scott, J. (1991). *Social network analysis.* London: Sage.

Spears, R., & Lea, M. (1992). Social influence and the influence of the 'social' in computer-mediated communication. In M. Lea (Ed.), *Contexts of computer-mediated communication* (pp. 30–65). Hemel Hempstead: Harvester.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in twitter events. *Journal of the American Society for Information Science and Technology, 62*, 406–418.

Trickett, A., Ellingworth, D., Hope, T., & Pease, K. (1995). Crime victimisation in the eighties: Changes in area and regional inequality. *British Journal of Criminology, 35*, 343–359.

Wilkinson, R., & Pickett, K. (2009). *The spirit level: Why equality is better for everyone*. Harmondsworth: Penguin.

Williams, M. (2006). *Virtually criminal: Crime, deviance & regulation online*. London: Routledge.

Williams, M. L., Edwards, A., Housley, W., Burnap, P., Rana, O., Avis, N., ... Sloan, L. (2013). Policing cyber-neighbourhoods: Tension monitoring and social media networks. *Policing & Society*. Advance Online Access.

Williams, M., Dicks, B., Coffey, A., & Mason, B. (2007). Qualitative data sharing and reuse: Mapping the ethical terrain. Methodological issues in data sharing and archiving: Briefing paper 2. Retrieved November 12,2012, from http://www.cf.ac.uk/socsi/hyper/QUADS/Briefing%20paper%20ethics.pdf