Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December 2014: 1–15

DOI: 10.1177/2053951714545135
bds.sagepub.com

SAGE

Big and broad social data and the sociological imagination: A collaborative response

William Housley[1], Rob Procter[2], Adam Edwards[1], Peter Burnap[1], Matthew Williams[1], Luke Sloan[1], Omer Rana[1], Jeffrey Morgan[1], Alex Voss[3] and Anita Greenhill[4]

Abstract

In this paper, we reflect on the disciplinary contours of contemporary sociology, and social science more generally, in the age of 'big and broad' social data. Our aim is to suggest how sociology and social sciences may respond to the challenges and opportunities presented by this 'data deluge' in ways that are innovative yet sensitive to the social and ethical life of data and methods. We begin by reviewing relevant contemporary methodological debates and consider how they relate to the emergence of big and broad social data as a product, reflexive artefact and organizational feature of emerging global digital society. We then explore the challenges and opportunities afforded to social science through the widespread adoption of a new generation of distributed, digital technologies and the gathering momentum of the open data movement, grounding our observations in the work of the Collaborative Online Social Media ObServatory (COSMOS) project. In conclusion, we argue that these challenges and opportunities motivate a renewed interest in the programme for a 'public sociology', characterized by the co-production of social scientific knowledge involving a broad range of actors and publics.

Keywords

Big Data, social media, COSMOS, public sociology, co-production, collaboration, methods innovation

Introduction

In this paper, we report on the work of the Collaborative Online Social Media ObServatory (COSMOS) project[1] as an evolving response to a number of fundamental methodological and disciplinary challenges for social science at the beginning of the 21st century. We explore the challenges presented to sociology and the social sciences in general as a consequence of the rise of commercial transactional data and the opportunities afforded by big and broad, publically available social media data for sociological and social scientific enquiry in the digital age. A key organizing principle here is the idea of collaborative observation. We frame these opportunities and challenges within the context of calls for a 'public sociology' (see Burawoy, 2005), whose aim is to establish a dialogue with a broad array of audiences beyond the academy and transform sociological practice. This dialogue requires an infrastructure for communication and collaboration to sustain it. As a contribution towards this, we present COSMOS, an open platform for social data analysis. Building and applying COSMOS necessitates engaging constructively with the 'computational turn' in sociology (also known as computational social science: for a critique, see Boyd and Crawford, 2012). COSMOS reflects how processes of social scientific knowledge production are adapting to meet the

[1]Cardiff University, Cardiff, UK
[2]Warwick University, Coventry, UK
[3]St Andrews University, Fife, UK
[4]Manchester University, Manchester, UK

Corresponding author:
Rob Procter, Warwick University, Gibbet Hill, Coventry CV4 7AL, UK.
Email: Rob.Procter@warwick.ac.uk

challenges and opportunities offered by new forms of social data. Of potentially even greater significance, we argue, is how COSMOS may contribute to the programme for a public sociology by providing a vehicle for the involvement of a broad array of publics in the co-production of social scientific knowledge.

Theoretical and methodological context

It is important to couch these challenges and opportunities within contemporary social thought. Many of these have been framed within the context of global complexity, mobilities and information flow (Urry, 2003), the rise of the networked society (Castells, 2011) and the consequential and constitutive effects of 'big and broad' social data upon social formations and relations (Ruppert et al., 2013). These represent profound questions for sociology as a viable empirical discipline that is able to speak truth to power while, at the same time, affording opportunities for innovation and re-invigoration in terms of theory, method, data and its relationship to society beyond the academy. The scale, complexity and speed of these transformations demand an interdisciplinary response, but they also speak to core sociological concerns that relate to classic questions of social organization, social change, and the integration and regulation of citizens within complex, late modern, globalizing, and interconnected social formations. However, these transformations raise questions about the capacity of academic social science to scope and make sense of them in comparison to other agents and institutions, where 'scoping the social' through access to 'big and broad' social data can help to realize competitive or strategic advantage for states, multinationals, and other agencies in the 'global race' in a 'runaway world' (Archibugi et al., 1998; Giddens, 2002). Furthermore, the theoretical consequences for reflexive modernization (Beck, 1992), liquid modernity (Bauman, 2000) and late modern social formation (Giddens, 2002) require further consideration in the light of the emerging contours of digital societies.

However, for sociology, and social science more generally, the present debate and response are centred on empirical concerns. The reasons for this can be understood to lie in the emergence of big and broad social data as a consequence of social and economic transformation realized through the digital revolution and rise of networked societies. Digital societies are self-referential, in the sense that they generate data as an accountable trace and functional pre-requisite for network and system integration. Furthermore, this data provides a powerful means of understanding and scoping populations and social life on a massive scale. This represents a challenge and an opportunity for sociology that is suffused with political, ethical and empirical issues. While all these are salient and mutually constitutive, it is questions of data and empirical enquiry that have brought these and related issues into sharp focus within sociology and the social sciences under the rubric of the 'social life of methods'; Ruppert et al. (2013: 24) state:

> ... we seek to unsettle debates about how the proliferation of the digital is implicated in large-scale social change and remaking the governance and organization of contemporary sociality (for instance, Castells' [1996] network society, or the notion of biopolitics ... we are concerned with the implications of digital devices and data for reassembling social science methods or what we call the social science apparatus. Here we build on our interest in elaborating the social life of methods ... through a specific concern with digital devices as increasingly the very stuff of social life in many locations that are reworking, mediating, mobilizing, materializing and intensifying social and other relations.

In their account of the 'coming crisis of empirical sociology', Savage and Burrows (2007) argue that, in previous 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 technologies, this claim has been compromised by the proliferation of transactional data generated, owned and increasingly analysed by large commercial organizations, as well as government departments. The availability to commercial enterprises of large volumes of continuously updated data on, for example, retail transactions, telephone communications, financial expenditure and insurance claims makes for an uncomfortable comparison with episodically generated datasets such as the census of populations,[2] general household surveys, police recorded crime, victim of crime surveys and labour market surveys on which academic sociology has traditionally relied, and provokes an existential question: is academic sociology "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 and Burrows, 2007: 886). Concerns about the marginality of academic sociology need revisiting in the light of the subsequent explosion of new digital 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'. Significantly, these technologies facilitate the mass communication and sharing of 'user-generated content', and have given rise to a form of mass, self-reported data about their users' daily routines, perceptions of, and sentiments

about, particular events. Twitter users, for example, post more than 500 million tweets per day; Facebook users post 9 million messages per hour.

Social and computational researchers have already begun to mine and 'repurpose' this naturally occurring, socially relevant data in their 'predictive' efforts. Tumasjan et al. (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 in 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 search engine with visits to the local doctor to epidemiologically trace the spread of the disease across the USA.[3] Another notable example is the wealth of social media communications about major incidents of civil unrest, such as the 'Arab Spring' (e.g. Howard et al., 2011; Stepanova, 2011) and the riots in English cities during August 2011 (Procter et al., 2013a, 2013b). These studies illustrate the potential significance of social media technologies for facilitating the harvesting 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 research methods (Cicourel, 1964). However, this proliferation of 'lively' social data poses a significant set of challenges that are still being confronted. Savage (2013: 4) states:

> My argument is that the 'Social Life of Methods' arises as part of a dual movement. These are, firstly, an increasing inter-disciplinary interest in making methods an object of study . . . I explore how this current poses a topical challenge to dominant instrumentalist readings of methods, which currently predominate in social science research. The second aspect of this interest is, however, less commented on, but in my view equally important. This is the crisis, increasingly evident in the 'research methods community' regarding positivist forms of knowledge, as forms of standardized data exceed the capacity of standard quantitative procedures to process and analyse them. The proliferation of 'lively' data has created an emergent space in which there is a dramatic potential to rethink our theoretical and methodological repertoires.

As academic researchers grapple with the methodological challenges posed by the growth of big and broad social data, it is important to acknowledge that these will not be resolved through internal dialogue alone. Big and broad social data raise significant ethical issues that demand an open debate with citizens about the role and status of academic research and the relationship between the academy and wider society. In his invitation for a 'public sociology', former President of the American Sociological Association Michael Burawoy identified the different ways in which social research is produced for and communicated to various publics, specifically students, policymakers and the broader citizenry (Burawoy, 2005). For Burawoy, a truly public sociology is one that sustains, nurtures and defends civil society against state and market pressures. However, in their paper on the 'coming crisis of empirical sociology', Savage and Burrows (2007) argue that the capacity of social research to realize such a public role is compromised by the emergence of big and broad social data streams generated by – and largely exclusive to – commercial transactional data. A consequence of this, in their view, is that commercial and private interests now have the capacity to envisage, indeed constitute, populations in powerful ways that are insulated from open and democratic scrutiny.

As we will argue, when set against the growth of openly available data from social media platforms and the momentum of the open data movement,[4] these threats to the legitimacy of academic social science may not be as grave as they seem. In our view, these developments provide an opportunity to forge a new relationship with society beyond the academy, which, if researchers are willing to seize it, may help to reinvigorate the programme for a public sociology.

Big and broad social data

The term 'big and broad' social data serves to draw attention to three salient dimensions that define new forms of social data: volume, variety and velocity, the latter reflecting its often real-time and rapidly changing character. The term has become one of the key phrases for describing the data deluge and the rise of digital infrastructure and device innovations that not only shape and constitute new forms of practice but also configure data streams in ways that reconfigure and constitute social relations and populations (Ruppert et al., 2013). Big Data is being generated in multiple and interconnecting disciplinary domains that include genetics, environmental science and astronomy, as well as within the social domain, where data is being produced through a myriad of transactions and interactions through multiple media and digital networks.

Technological innovation in digital communications, epitomized in the shift from the informational web (Web 1.0) to the interactional web (Web 2.0), provokes

new opportunities and challenges for social research. Web 2.0 technologies, particularly the new social media platforms (e.g. social networking, blogging and micro-blogging), as well as the increased accessibility of the Web through portable and ubiquitous devices like smartphones, tablets and net books generate new forms of data which are of significance for social research, as well as stimulating the development of new methods and techniques for analysis. At the same time, the increasing adoption of open data principles by large public bodies in the UK and elsewhere is giving researchers fresh opportunities to interrogate new digital data streams to answer social science questions.

Large national and multinational corporations have recognized the power of Big Data to help them spot business trends; they have developed infrastructure and strategies to collect a wide range of data, and concerns have grown that traditional social science research methods would not be able to compete. It is feared that big business and other organizations with deep pockets can use the data they gather to group people together into populations that are new and powerful, but are inaccessible to public social science or, indeed, to any meaningful public scrutiny. This has led to the so-called 'empirical crisis' identified above. However, the rise of digital innovations characterized by interaction, participation and the 'social' provides an opportunity to explore ways in which this asymmetrical relationship with data and analytic capacity might be confronted. In particular, they offer the prospect of new ways of engaging with diverse publics, such as 'citizen social science' where members of the public can assist with research through crowdsourced coding and registration of their beliefs and opinions at volume in relation to key sociological concerns (see Procter et al., 2013c). We see these developments as affording a way forward for confronting the empirical crisis within the social sciences through exploiting these new forms of open data and for developing an additional resource for a public sociology that has citizen participation at its core. Realizing this transformation entails the development of a participatory infrastructure of digital 'observatories' and 'collaboratories'. To this end, we present the COSMOS project as a potential exemplar and early prototype.

Challenges, opportunities and digitally re-mastering the classic questions

In contrast to the pessimism of the Savage and Burrows thesis, we see opportunities as well as the challenges presented to social science by innovations in digital technologies. Broadly defined, the evolving field of 'digital social research'[5] has begun to recognize the value of big and broad social data. The COSMOS project forms a part of this evolving research field and is itself the product of intensive inter-disciplinary collaboration between the social scientists and computer scientists co-authoring this paper. The computational engineering involved in the COSMOS project is discussed in further detail below. In this section, our focus is upon the social scientific relevance of digital technologies and the kind of data they produce. In brief, we want to argue that far from undermining the social scientific programme pursued in the latter half of the 20th century and epitomized in C. Wright Mills' (1959) vision of the sociological imagination, these technologies and their allied data have the potential to 'digitally re-master' classic questions about social organization, social change and the derivation of identity from collective life.

Now that people have enthusiastically adopted social networking platforms right across society, and mobile devices such as smartphones and tablets are routinely used to access information and to interact with friends and with strangers alike, new forms of data are being created that are highly significant for social research. 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 (Edwards et al., 2013). Some commentators suggest that 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 such as parenting skills or medical self-diagnosis. It can also be argued that digital social research *augments*, but needs to be used in conjunction with more traditional methods. C. Wright Mills' identification of three classic questions that underpin the sociological imagination are useful for clarifying the distinctive contribution of digital social research: what can it do that traditional methods cannot in understanding how social organization and relations are constituted, how do these change over time and how do they generate social identities? It is argued that digital social research, particularly in the context of 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 unrealizable possibility of studying social processes as they unfold at the level of populations as contrasted with their official construction through the use of 'conventional' research instruments and curated datasets (see Table 1 for a summary).

An exemplar of this is the possibility of augmenting traditional methods of psephology by registering voting

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

		Research data/design	
		Locomotive	Punctiform
Research strategy	Intensive	E.g. ethnography/observational studies	E.g. Cross-sectional qualitative interviewing
	Extensive	E.g. New social media analysis: population level, naturally occurring data in real/useful time	E.g. surveys: (cross-sectional, longitudinal); experimental studies

Source: Edwards et al. (2013: 248).

sentiments expressed through micro-blogging or traditional methods of urban political analysis by investigating the role of social media in shifting local policy agendas (e.g. Bonsón et al., 2012; Lidén and Nygren, 2013). In these two examples, new social media can be seen to have the potential both to re-organize and change social relations, while leaving a digital footprint that can be collected, analysed and visualized. The pace at which this footprint is accumulating and the recursive qualities of social media also have the potential to re-orient social research. An early example of this was the role of social media in propagating rumours about riotous activity in English cities in August 2011 and in dispelling them (Procter et al., 2011, 2013a). There are also examples of the role of social media in propagating not only hateful sentiments but also counter-speech, which challenges bigotry and other misinformation. An instance of this is the recent micro-blogging reaction to a UK television documentary ('Benefits Street'), which followed the lives of welfare benefits claimants living on a street in the city of Birmingham in the English Midlands. The Twitter timeline for #benefitsstreet demonstrates the pace at which misunderstanding about the numbers and characteristics of benefits claimants can be challenged both by individual micro-bloggers and by those representing campaigning groups and trade unions.

As indicated in Table 1, the distinctive contribution of big and broad social data such as social media, as both a subject and means of social research, can be clarified through reference to concepts of research strategy and design in the philosophy of social science (e.g. Edwards et al., 2013; Sayer, 1992). In this literature, the distinction between intensive and extensive research is used to differentiate research strategies that are concerned with investigating how processes work in a particular case from those concerned with identifying the 'regularities, common patterns and distinguishing features of a population' of cases (Sayer, 1992: 243). Another useful distinction is between research designs that seek to capture the locomotion of social life, the idea that social relations always have to be accomplished and are therefore subject to reformation if not transformation, and those that seek to puncture this process at certain points in order to capture a 'snapshot' of how social relations are configured at any one moment. The former designs imply research that can support the continuous observation of social life, which, prior to the advent of big and broad social data, necessitated forms of qualitative inquiry and ethnographic immersion in the social process in question and this limited observation to the study of 'individual agents in their causal contexts' (Sayer, 1992: 243). By contrast, the distinctive quality of big and broad social data for research is the possibilities it provides for the continuous ('real-time') observation of populations hitherto only accessible through episodic and retrospective snapshots gleaned through such instruments as household surveys and census data, longitudinal studies of cohorts and experiments measuring pre-test and post-test conditions. In these terms, the distinctiveness of big and broad social data is the possibility of extensive research into the locomotion of social life, such as the unfolding of election campaigns, the shaping of policy agendas in local government, the prevalence of suicidal ideation, the propagation of bigoted and prejudicial opinions and the 'sensing' of crime. As emphasized elsewhere (Edwards et al., 2013), however, the real transformative power of big and broad social data is in its use to augment and re-orientate rather than replace the other more established research strategies and designs depicted in Table 1.

The COSMOS project as a response

The COSMOS project represents an attempt to forge interdisciplinary working between social, computing and computational scientists as a means of realizing the theoretical, methodological, empirical and public objectives identified above. A genuine conversation and orientation towards interdisciplinarity is key to responding to the challenges of big and broad social

data and the emerging architecture of digital societies. In a recent article, Tinati et al. (2013: 175) state:

> Unless sociologists are prepared (and able) to acquire sophisticated computational expertise, we must collaborate with computer scientists . . . to develop multidisciplinary curricula and research that transcend the usual disciplinary boundaries. We have experienced first-hand the challenges arising from the different epistemologies, histories and languages of sociology and computer science, which raise questions about the wider politics of knowledge and dynamics of power and identity that arise in multidisciplinary work.

Practices associated with collaborative working and teamwork are important for realizing interdisciplinary work of this sort. While not a focus of this paper, the role of interdisciplinary work in networked and distributed teams and citizen research is worthy of future scrutiny and consideration (Dutton and Jeffreys, 2010). Furthermore, it is important to note that the co-production of digital tools has to sit side by side with theoretical and methodological concerns. To this extent the COSMOS platform is merely one expression of a wider programme of research within a collaborative observatory framework where other 'offline' research methods are also of importance – not least in relation to the ongoing refinement of algorithms via expert and lay input through a process of 'collaborative algorithm design' (Edwards et al., 2013: 256–257). For the remainder of this paper we will focus on the features of the COSMOS platform and consider how it links to a public sociology agenda and the challenges and opportunities outlined earlier.

The COSMOS platform

The COSMOS platform provides an integrated suite of computational tools for harvesting, archiving, analysing and visualizing social media data streams using publicly accessible application programming interfaces (APIs). In this paper, we focus on Twitter data as it arguably provides the most open and voluminous social media data source and has thus become established as a key data source for public opinion and behaviour mining. Twitter data has been used to measure public mood (Bollen et al., 2009), opinion (Pak and Paroubek, 2010; Thelwall et al., 2011), tension and cohesion (Burnap et al., 2013a; Williams et al., 2013) and to explore communication patterns (Bruns and Stieglitz, 2012).

The COSMOS platform currently provides nine modes of analysis, some of which operate at the individual tweet level and others at a corpus level (i.e. tweet collections). These can be applied individually or in combination to enable the exploration of, and 'drilling down' into, datasets as a precursor to more detailed interrogation.

Individual tweet level

- Gender identification is used to derive the portrayed gender of the person who posted the tweet (the tweeter).
- Language detection is used to determine the language used in the text of the tweet.
- Sentiment analysis is a form of opinion mining that attempts to derive a score (positive or negative) to measure the polarity and strength of mood expressed in a tweet (Thelwall et al., 2011).
- Tension detection was developed specifically for COSMOS. It implements a conversation analytic method – membership categorization analysis – combined with lexicons of expletive terms, tension-specific degradation terms and attribution terms to classify tweets on a three-point ordinal tension scale (Burnap et al., 2013a).
- Geo-spatial location assigns a sending location to the tweet. A small proportion of tweets (~1%) currently have global positioning system (GPS) co-ordinates included within their metadata. For those that do not, this tool attempts to derive a probable location from user profile metadata and keyword matching of text referring to place.

Corpus level

- Keyword frequency analysis visualizes occurrences of specified keywords as a bar chart over time. This allows the researcher to identify visually points of high and low activity in relation to an event or topic. COSMOS visualizes frequency using three units of time – by day, hour and minute – each visualized on its own timeline (see Figure 1).
- Social network analysis enables visualization of the interactional relationships between groups of Twitter users (see Figure 4 for an example).
- Qualitative overview provides a list of the text in all tweets. This can comprise all tweets within a specified time range, tweets that match the parameters identified using the filters, or a combination of both. The text of each tweet is displayed, along with two attributional annotations: the gender of the tweeter and the sentiment scores (positive and negative) calculated based on tweet content. This gives the researcher an 'at a glance' view of the filtered dataset, which can support identifying key topics, events, opinions and perspectives from the text.

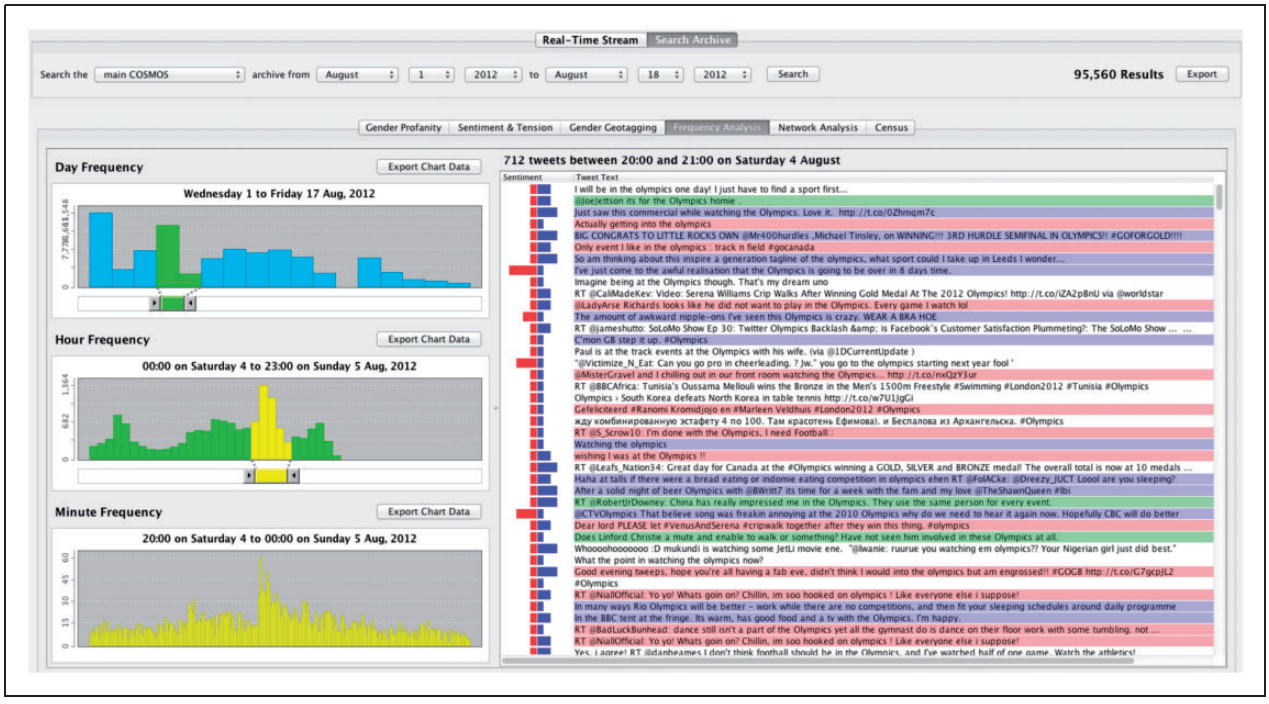

Figure 1. COSMOS frequency analysis.

For a detailed description of the COSMOS platform and its tools, see Burnap et al. (2014a).

Linked data

The COSMOS platform was conceived from the beginning as providing ways to link social media data with other sources of social data, including key socio-demographic datasets. At the time of writing, the platform has access to the UK Police API, which provides crime data on a local district level for the previous month, and is in the process of establishing access to the UK Office for National Statistics, which holds census data (as well as many other datasets), including, *inter alia*, district-level unemployment, ethnic composition and population size.

One way in which COSMOS supports data linking is through geography (see Figure 2).

Summary

The COSMOS platform is currently undergoing beta testing and additional analytical tools are in development. Because of the inter-disciplinary make-up of the project team and the core role played by its social scientist members, development has always been driven by an evolving understanding of how computational methods can best serve the needs of social research. The guiding principle has been to explore ways in which computational social science can make analysis of big and broad social data tractable for the established study principles of qualitative and quantitative research, while creating the space for methodological innovation.

Overview of current research

COSMOS is based on interdisciplinary, collaborative working where a combination of theory, method and data informs our empirical research. In this section, we present four examples of current research in order to illustrate the potential for collaborative observatories to generate empirically and theoretically informed insight into the use of social media-as-data within sociological and social scientific research.

In a recent paper, Tinati et al. (2013: 2) argue that sociologists have been slow to respond to the challenges of big and broad social data and some of the opportunities afforded by social media. They state:

> . . . to date, the scope for pushing this research forward has been methodologically limited because social scientists have approached Big Data with methods that cannot explore many of the particular qualities that make it so appealing to use: that is, the scale, proportionality, dynamism and relationality described above. Rather, Big Data has commonly been approached with small-scale content analysis – looking at small numbers of users – or larger scale random or purposive samples

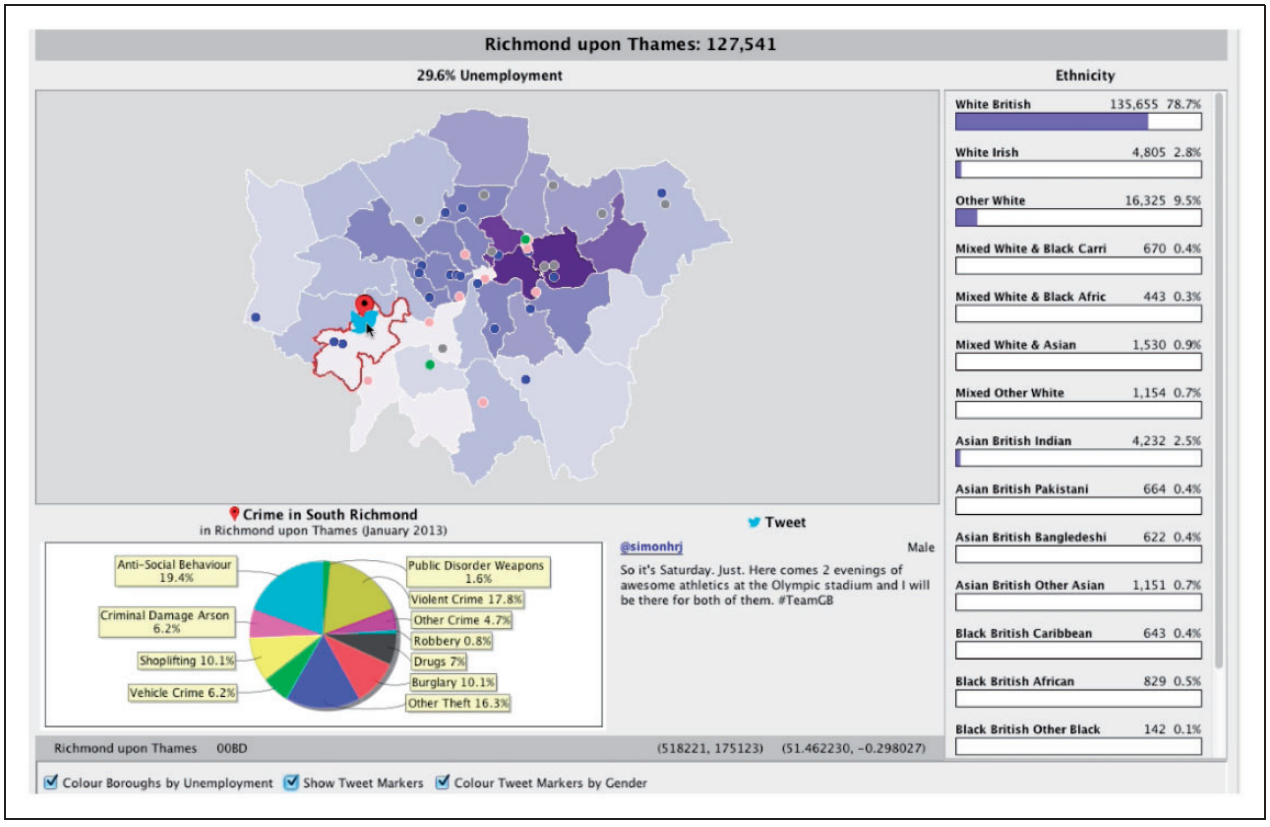

Figure 2. Linking social media data and administrative datasets through geography.

of tweets. Rendering Big Data manageable in this way overrides its nature as 'big' data, bypassing the scale of the data for its availability or imposing an external structure by sampling users or tweets according to a priori criteria, external to the data themselves. Furthermore, most previous social science studies are snapshots, categorising content and user-types rather than following the data as it emerges dynamically or exploring the nature of the social networks that constitute Twitter.

The body of work now emerging out of the COSMOS project (Burnap et al., 2013a, 2013b, 2014b; Edwards et al., 2013; Housley et al, 2013; Procter et al., 2013a, 2013b, 2014; Sloan et al., 2013; Williams et al., 2013) exemplifies how the concerns raised by Tinati et al. may be addressed. This work reports extensively on interdisciplinary collaborative platform and tool development, methodological issues in social media analysis, the use of social media analytics in the study of contemporary social phenomena, and a consideration of wider methodological and theoretical issues for sociology and social science. Space prevents a recapitulation of the points here. However, it is worth reporting some observations derived from our current projects at this point.

Social media, demographic proxies and crime sensing

A fundamental problem for researchers is that tweets are 'data-light', i.e. they lack important demographic data, e.g. gender, location, class and age, about their users (Gayo-Avello, 2012; Mislove et al., 2011). Yet although such data is not present in an explicit manner, the tools available on the COSMOS platform enable it to be inferred with a relatively high level of confidence (Sloan et al., 2013). These derived metadata are automatically generated and added to harvested tweets.

The London 2012 Olympics provide an example of how demographics can aid interpretation of social media data. Figure 3 is a sentiment graph covering between 20:00 and 23:00 on Saturday, 4 August 2012 (taken from Burnap et al., 2013). This date is more commonly known as 'Super Saturday' as it was the evening during which Team GB won three gold medals. The data used to produce this graph consists of tweets containing the hashtag #TeamGB.

Looking at the top two lines we can see that the major peaks in positive sentiment for Mo Farah's progress in the 10,000 m and the moment when Jessica Ennis captures the gold in the heptathlon are female dominated, i.e. female tweeters show higher levels of positive sentiment than male tweeters. Observations such as this

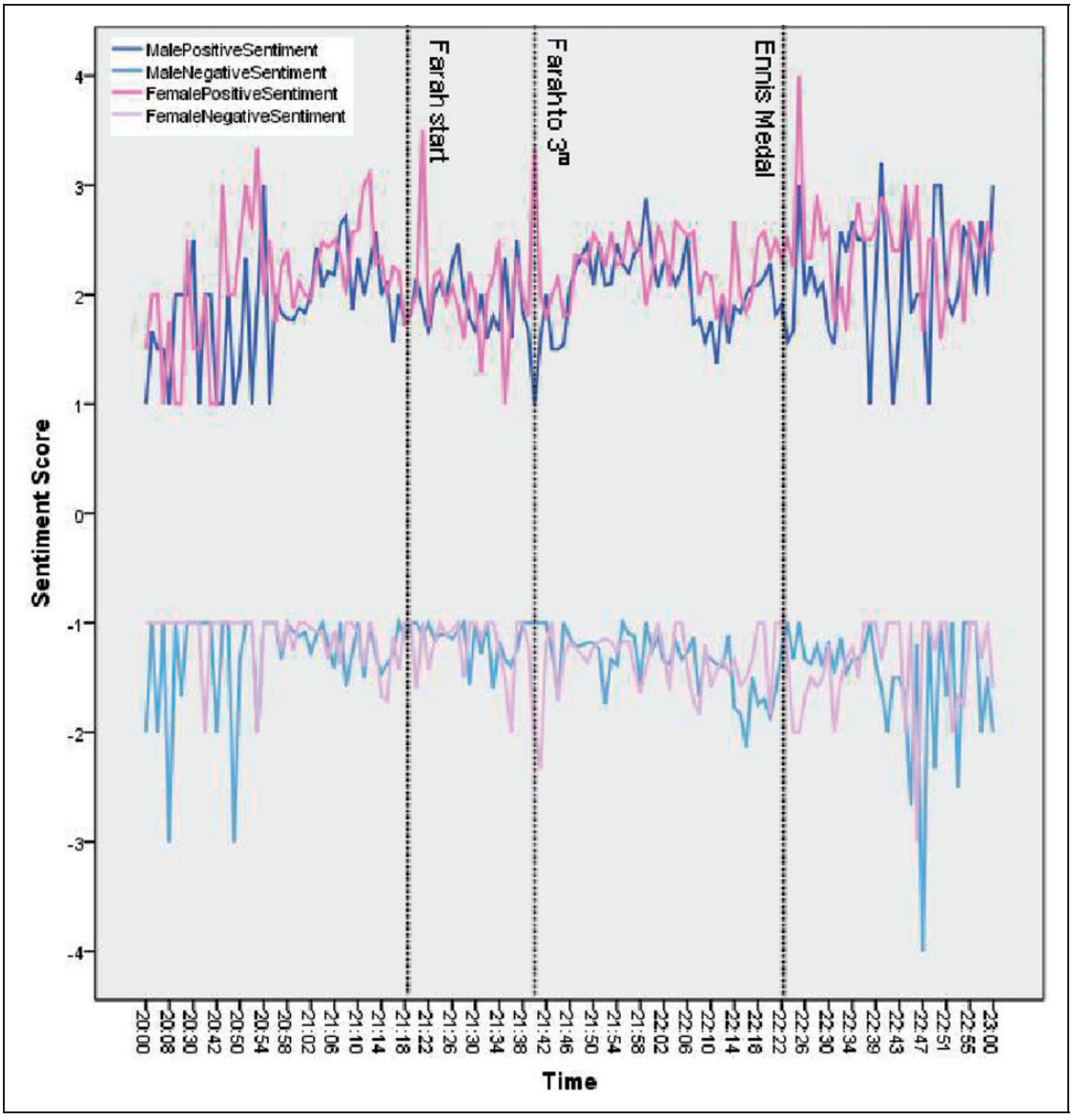

Figure 3. Male/female sentiment of tweets containing #TeamGB between 20:00 and 23:00 on 4 August 2012 (Burnap et al., 2013).

generate new questions over how people engage with social media and the influence of gender on content.

One current COSMOS project is using census, crime and tweets to explore whether crime can be 'sensed' through social data via the signatures (social media) and context (area demographics) of real world events.[6] This is an example of how linked social media and curated data can be used to enrich statistical models of social phenomena in ways that may be able to account for complex temporal and spatial factors.

Understanding social media use at the local and civic level

We are conducting an ongoing study of Twitter use in two districts in the UK cities of Cardiff and Manchester to compare emerging trends in the use of this social media at the local and civic level and the extent to which social media offers ways for the re-shaping of citizen engagement in civil society (Procter et al., 2014). Methodologically, our goal is to explore how the analytical tools and capabilities of the COSMOS platform can be applied to scope big and broad Twitter data streams to a local or civic level and to assess the sociological benefit of doing so in terms of the capacity to generate some form of 'scopic' sociological insight. It illustrates how COSMOS supports the combination of computational social science methods applied to big social data (and social network analysis in particular) with in-depth, qualitative analysis.

We are interested in looking for evidence of the impact of early adopters and 'innovation intermediaries'

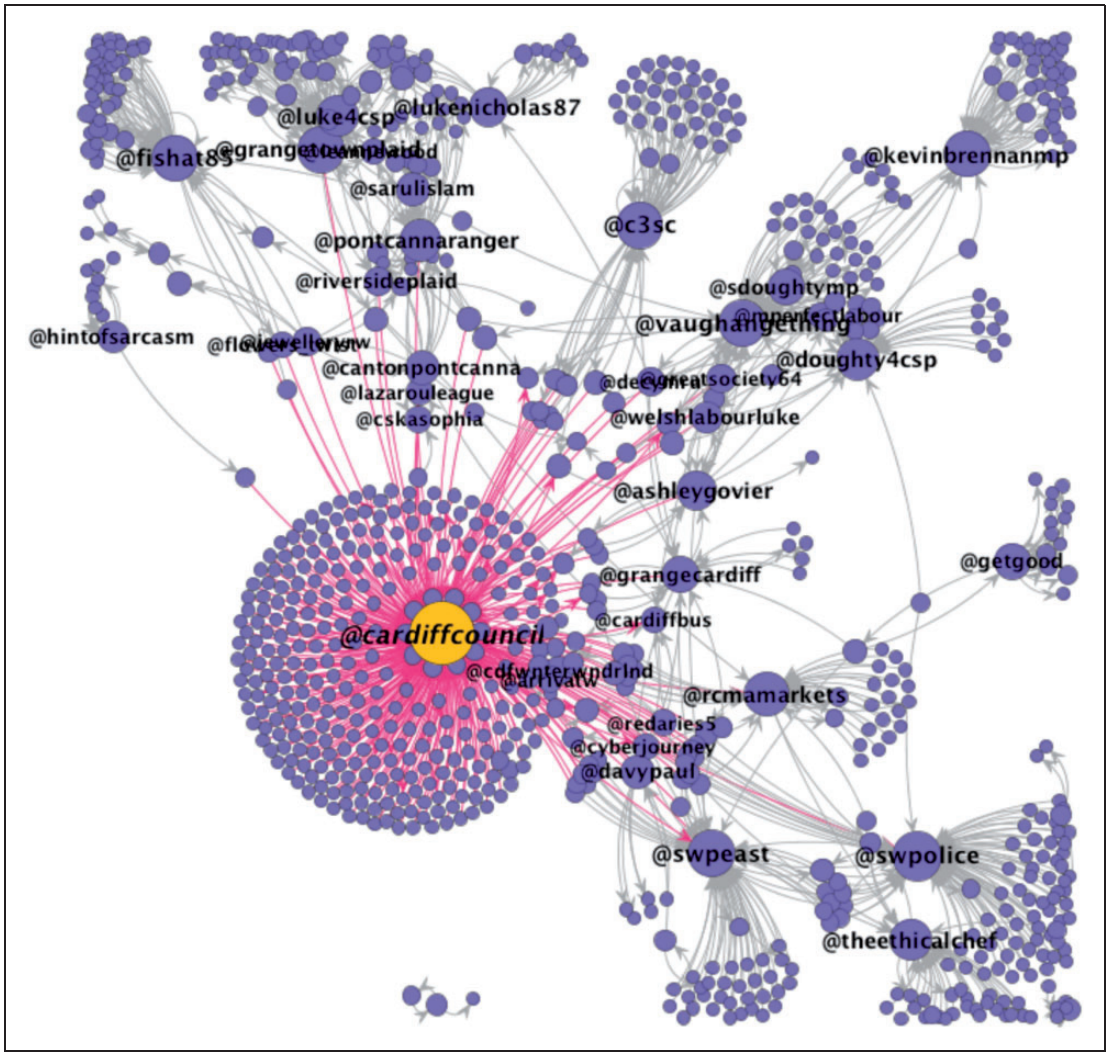

Figure 4. Visualization of the social network of Twitter communications in West Cardiff.

(Williams et al., 2005), that is, individuals and groups with the skills and resources to shape the use of social media as a tool for civic participation and how they orientate to achieving their objectives. In the UK, for example, third sector non-governmental organizations (NGOs) such as MySociety ("using digital technologies to make people powerful") are noteworthy for their use of Web 2.0 technologies to gather opinions of citizens and act as advocates for local causes.

The study is based on tweets harvested from the Twitter streaming API between 25 October 2012 and 14 January 2013 containing hashtags and accounts that we identified through an initial snowballing exercise as being associated with the respective communities. This resulted in a total collection of over 100,000 tweets and re-tweets. Applying social network analysis to this sample enables us to identify accounts from which tweets and re-tweets originated. In particular, we are interested in three specific measures (see Figure 4):

1. In-degree (those accounts that were the most targeted by other tweeters and re-tweets);
2. Out-degree (those accounts that posted the most tweets and whose tweets were the most re-tweeted);
3. Between-ness centrality (those accounts that ranked highest in terms of receiving and posting tweets).

Using these measures and more general measures of activity on Twitter, we identified 50 accounts for further investigation, i.e. qualitative analysis of Twitter postings to determine their topics and interviews to explore the variety of strategies and tactics deployed in their use of social media within the local civic sphere. Each of these interviews was transcribed and coded using the NVivo software package.

Social network analysis of the dataset reveals that the local state was overwhelmingly represented with the local council and police force figuring prominently

and representing key nodes in the network. Nevertheless, it also reveals the presence of several NGOs acting as 'bridges' between local state actors and individual citizens.

One local civil society organization explained how they had used social media to mount a successful campaign against a council scheme:

> When [council xxxx] proposed introducing residents' parking in [yyyy], we created a separate Twitter account called ''Saveyyyy'' and garnered peoples' opinions on the issue. Enabled us to effectively mobilise a campaign to object to the scheme...

Another interviewee explained:

> I have been doing Twitter to communicate with third sector organizations in [xxxx], but particularly our member organisation... that you mentioned, retweeting information and getting the message out, shoring up relationships. That sort of thing...
> I guess it is more to increase the following, but we want to represent the third sector organisations in [xxxx], so we have got our member organisations and we are looking at increasing our membership, so it's to engage with our member organisations, but also hopefully other third sector organisations in [xxxx] who aren't members and maybe haven't heard of us and who will think gosh, they have got a lot of resources or information.

So far, our study provides mixed evidence for the proposition that social media is enabling a radical reshaping of local civil society. Powerful local state actors and established political groupings have been amongst the most enthusiastic and effective early adopters of new digital communications technologies, but there is also evidence of local civil society organizations using social media effectively to promote their agendas.

These results are only a snapshot of social media use as a tool for civil society promotion and the relatively short sampling period may bias the results, which motivates continuing research into how early adopters are using these technologies and how this is influencing their potential for radically re-shaping citizen engagement in local civil society.

'Hate' speech and social media: Understanding users, networks and information flows

The rapid and widespread uptake of social media platforms brings both benefits and risks for civil society and new challenges for agencies responsible for ensuring that the boundaries of acceptable and legal behaviour are not crossed. In this respect, the proliferation of the so-called 'hate speech' in social media is an area of growing concern, as recent high-profile examples confirm.[7] The most senior prosecutor in England and Wales recently acknowledged the harm that can be caused by hate speech on social media and explained that ''banter, jokes and offensive comment are commonplace and often spontaneous'' and ''communications intended for a few may reach millions.''[8] This project is a study of the migration of hate speech to social media platforms and focuses on understanding the propagation of this type of antagonistic language.[9]

The project poses several key questions: (i) can we identify hateful and antagonistic social media content, as well as attempts to counter it, in terms of key events, linguistic characteristics, sentiment and tension? (ii) Can we profile hateful and antagonistic social media networks in relation to user behaviour and interaction, building on the previous question to develop a typology of users? (iii) Can we triangulate the above analysis with other forms of open data, such as the new Google Trends[10] metrics to validate the propagation of hateful content into online environments beyond social networks? (iv) Can we utilize the data derived from the above questions to build probabilistic models to forecast the emergence and evolution of information flows within social media networks through which hate-related content is transmitted? And (v) can the model and methodology inform the social scientific interpretation of how hateful content travels and is impeded online, drawing on social scientific concepts such as responsibilization (Garland, 2001) and nodal governance (Shearing and Wood, 2007) as framing devices?

To date we have generated hate speech corpora covering content that is considered homophobic, racist, sexist and disablist. These datasets are being examined using various statistical modelling techniques to identify enablers and inhibitors to hate speech propagation (Burnap et al., 2014b). The significant covariates of propagation can be used by regulatory authorities to potentially stem the spread of hate speech in the social media eco-system. Our most recent results have shown that racial tension on Twitter can propagate around major sporting events and can be identified using bespoke tools such as the COSMOS tension engine that is usable by law enforcement to help inform operational decisions (Williams et al., 2013).

Citizen social science

We are committed to developing ways in which the COSMOS platform can be used to facilitate public participation in social science. One approach we are

currently exploring is 'citizen social science', where members of the public can assist with research, and record their beliefs and opinions at volume (Procter et al., 2013c).

Our interest in encouraging citizen social science has the very pragmatic goal of securing scalable human effort for the analysis of large social media datasets, as projects such as Galaxy Zoo[11] have already demonstrated for the physical sciences. However, we argue this may be a potentially significant step towards realizing the programme for a public sociology. It seems to us that an important principle for motivating volunteer effort is offering meaningful engagement with the research. At a minimum this might involve providing volunteers with access to the results made possible by their efforts. More ambitiously, we see citizen social science as providing a basis for forging a new relationship between the social science academy and society.

> Huge potential exists to harness the power of crowdsourcing for the study of society and human behaviors . . . but it's just not happening as well as it could . . . it seems odd that social science researchers appear to have been comparatively slow to investigate the potential of crowdsourcing . . . social research could be enhanced by the involvement of the public – from helping to set research agendas, contributing to and helping to analyse data sets, to formalising findings and conclusions. Social science issues are human issues, after all – they are about how we relate to each other and organise our society and economy – so there seems to be a natural fit with crowdsourcing that's largely being overlooked. This raises some obvious and legitimate concerns – from representation to research ethics and integrity – but none of these seem insurmountable. Indeed, social scientists would surely benefit from greater public engagement with their work. The prize is surely quicker, cheaper and more imaginative research – the findings from which could benefit us all. (Harris, 2012)

Our aim over the long term is to develop the COSMOS platform as a 'collaboratory', an element of a participatory research infrastructure supporting public engagement in a range of activities that includes the exchange of ideas, debates about the shape of institutions, current social problems, opportunities and events, as well as the co-production of social scientific knowledge through citizen social science, where publics act as vital sensors and interpreters of social life. However, any emerging citizen social science will also have to take account of other relational elements and configurations that include social class, race, gender, sexual orientation and geography, in addition to constellations of expertise and broader common sense understanding. The synthesis of crowdsourcing techniques with a sociologically informed citizen social science remains public sociological work in progress.

Concluding remarks

Developments linked to the emergence of big and broad social data are happening rapidly, and we cannot be certain what impact it will have on research processes. It is possible that it will promote the use of computational social science methods in place of more traditional quantitative and qualitative research methods. It may also influence thinking and re-orientate social research around new objects, populations and techniques. However, we think it is most desirable that new methods be used in conjunction with the existing ones, to make research richer and more nuanced, and we have attempted to motivate this synthesis through examples of our current research summarized above. The analysis of social processes as they actually happen is bound to give researchers insights and interesting avenues to explore that are absent from the official construction of events that is available via traditional research instruments and curated datasets. The COSMOS platform has been developed to help academic researchers embrace this opportunity.

This is not without its challenges, however. An initial hypothesis of ours was that the high volume and velocity of social media communications, as a form of big social data that is user-generated, would enable us to better access or 'sense' civil society without the interlocution of administrative or professional categories. However, as our discussion of initial findings suggests, administrative and professional organizations have been amongst the most enthusiastic early adopters of social media communications and thus the next phase of social media analysis will need more refined methods for differentiating between the kinds of actor generating this 'Big Data', including, of course, the impact of 'botnets'. An obvious example of this is the use of botnets to re-tweet and propagate campaigning materials during elections. As a consequence, this next phase will also have to develop methods for understanding the recursive qualities of social media communication and whether it is possible to disambiguate types of human and non-human actor in social media communications and, in turn, the consequences of their interaction for shaping social relations, such as the outcome of election campaigns. COSMOS has made a start on this kind of analysis through its interrogation of information flows and what they tell us about the patterns of human interventions in Big Data, such as the authoritative rebuttal of rumour, prejudice and bigotry.

Returning to our earlier distinction between the qualities of social media as both a means for, and subject

of, social science, it is possible to identify a number of challenges for augmenting and re-orienting social research through use of the Big Data generated by social media. First, with respect to augmentation, there is a need to reconstitute the 'Big Data' generated by social media into units of analysis that enable it to be linked to datasets held by administrative, professional and commercial organizations. As discussed above, COSMOS has made some headway in this by examining how the sensing of crime through social media can be meaningfully contrasted with police-recorded crime rates. The granularity of insights into the pattern of social relations that can be gleaned through linking Big Data with other sources of data also needs to address the wealth of administrative and curated datasets held by local authorities, not just those stored in national archives. As we noted earlier, here the increasing adoption of open data principles by public bodies gives grounds for optimism.

Second, with respect to re-orientation, there is the issue of access to social media data, and here the outlook is less clear. Although free, open access to social media datasets is subject to constraints, as companies seek to monetize their data assets, it is nevertheless currently possible for academics to harvest significant and useful volumes of data at no cost. This present arrangement is not sustainable for three reasons, however. First, it does not meet the needs of all researchers: inevitably, some research will require more data than is available without cost and charges imposed by social data resellers[12] that are often too expensive for most academic researchers. While it has been possible for some researchers to negotiate individual deals with social media companies, this solution is unsatisfactory for obvious reasons: it does not scale and benefits the few at the expense of the majority. Second, it leaves researchers at the mercy of data providers' terms and conditions, which may change at any time. Third, where these terms and conditions prohibit sharing of data, they actively inhibit the capacity of the research community to test and validate findings, a cornerstone of empirical research practice and trust in scientific knowledge production.

Resolving these tensions calls for concerted action by research agencies and other stakeholders to negotiate with social media platforms not-for-profit access, under suitable terms and conditions, to social media datasets at no charge.[13] Indeed, there are some grounds for optimism. In the USA, for example, the Library of Congress announced in 2010 that it had reached agreement with Twitter on the archiving of all public tweets, with the promise that the archive will be made available to researchers.[14] At the time of writing, the archive remains inaccessible. Nor, of course, does the Library of Congress plan cover other social media.

Our aim in the COSMOS project is to help confront the challenges to the social sciences that have been raised by Burawoy, Savage and Burrows and many others. We are doing it by engaging with new forms of social data and developing in COSMOS a resource for public sociology that has citizen participation at its heart. Social science is by no means the only discipline in which knowledge production is changing and becoming more 'public': but it is probably the one in which the change seems especially appropriate. Now research can start to be done differently, and communicated differently. These changes will inevitably force us to rethink the role of the academic social scientist in the future. One way forward would be for academic social scientists to actively seek collaborations with groups, both professional and lay, involved in doing various kinds of 'practical sociology'. Examples of the former might include journalists[15] who increasingly find themselves needing to analyse large datasets in order to report news stories;[16] examples of the latter might include community activists who wish to engage with policymakers over issues of concern. As academic social scientists, we are intrigued by the prospects of emulating the example of voluntary organizations such as the Public Laboratory for Open Technology and Science,[17] which seek to promote the transfer of skills and technologies for environmental science to community groups.

Big and broad social data has given fresh stimulus to debates about research ethics (see e.g. Boyd and Crawford, 2012), much of which focuses on the issue of people's right to privacy. At one level, we would argue that questions about the 'public' or 'private' character of the communications captured as big social data, its panoptic use for mass surveillance and its synoptic use for challenging elite constructions of social problems and so forth, ought themselves to be the subject of ongoing deliberation and empirical inquiry as part of the search for a consensus. We would also argue that we must not lose sight of the broader issue of the ethics of research and innovation (see e.g. Stahl et al., 2012). We see the promotion of a public sociology as an important step towards both of these objectives.

Finally, and following on from the above, the boundaries of social science research practice are becoming more porous. As with other disciplines, social scientific knowledge production is changing, potentially becoming more 'public', the emergence of citizen social science being a case in point, but also in terms of the ways in which research is communicated and the rise of the 'networked researcher'. These developments require a rethinking of the role of the academic social scientist. The opportunity exists for reinvigorating the programme for a 'public sociology'.

Taking this opportunity involves embracing openness and public dialogue in a digital age, and having the capacity to engage in timely ways – including in 'real-time' – with unfolding events and social problems as they emerge.

Acknowledgements

We wish to thank the UK Economic and Social Research Council (grant numbers ES/K008013/1 and ES/J009903/1), the National Centre for Research Methods, the Digital Social Research programme and the UK Joint Information Systems Committee (Digital Infrastructure Research Tools Porgramme) for funding this work.

Declaration of conflicting interest

The authors declare that there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. www.cosmosproject.net
2. The UK decennial census is now recognized by the Office of National Statistics as no longer being fit for purpose in its current form.
3. The predictive accuracy of Google flu trends has subsequently been called into question. See www.theguardian.com/technology/2014/mar/27/google-flu-trends-predicting-flu
4. Progress towards open data is subject to national differences. For UK developments in this arena, see data.gov.uk/
5. http://www.esrc.ac.uk/research/research-methods/dsr.aspx
6. http://www.esrc.ac.uk/my-esrc/grants/ES.J009903.1/outputs/read/69ae0566-fa1a-4150-83fd-24293e73e505
7. For example, http://www.guardian.co.uk/uk/2012/jun/26/police-alleged-racist-abuse-twitter and http://www.guardian.co.uk/uk/2012/may/22/muamba-twitter-abuse-student-sorry
8. http://www.bbc.co.uk/news/uk-19660415
9. http://www.esrc.ac.uk/my-esrc/grants/ES.K008013.1/read
10. Google Insights for Search was recently incorporated into Google Trends. It is not yet clear if the API provided for the former service will remain accessible under the new arrangements.
11. http://www.galaxyzoo.org/
12. E.g. Gnip, DataSift.
13. In February 2014, Twitter announced a 'data grants initiative'. In response, over 1300 proposals were received; six were selected.
14. http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/
15. See, for example, the 'reading the riots' project, Lewis et al. (2011).
16. This has given rise to the new specialism of 'data journalism'. News media organizations have also been at the forefront of experiments in citizen journalism and crowdsourcing data analysis. For an example of the latter, see http://www.theguardian.com/news/datablog/2009/jun/18/mps-expenses-houseofcommons
17. publiclab.org

References

Archibugi D, Held D and Köhler M (eds) (1998) *Re-imagining Political Community: Studies in Cosmopolitan Democracy*. Stanford, CA: Stanford University Press.

Asur S and Huberman BA (2010) Predicting the future with social media. In: *Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Toronto, 2010 IEEE/WIC/ACM international conference. IEEE, Vol. 1, pp. 492–499.

Bauman Z (2000) *Liquid Modernity*. New York: John Wiley & Sons.

Beck U (1992) *Risk Society: Towards A New Modernity*. London: Sage.

Bollen J, Pepe A and Mao H (2009) Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: *Fifth international AAAI conference on weblogs and social media (ICWSM)*, Barcelona, 2009.

Bonsón E, Torres L, Royo S, et al. (2012) Local e-government 2.0: Social media and corporate transparency in municipalities. *Government Information Quarterly* 29(2): 123–132.

boyd d and Crawford K (2012) Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662–679.

Bruns A and Stieglitz S (2012) Quantitative approaches to comparing communication patterns on Twitter. *Journal of Technology in Human Services* 30(3–4): 160–185.

Burawoy M (2005) For public sociology. *American Sociological Review* 70(1): 4–28.

Burnap P, Rana OF, Avis N, et al. (2013a) Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change*.

Burnap P, Avis NJ and Rana OF (2013b, in press) Making sense of self-reported socially significant data using computational methods. *International Journal of Social Research Methodology* 16(3): 215–230.

Burnap P, Rana O, Williams M, et al. (2014a) COSMOS: Towards an integrated and scalable service for analysing social media on demand. *International Journal of Parallel, Emergent and Distributed Systems*. Epub ahead of print, 1–21.

Burnap P, Williams ML, Sloan L, et al. (2014b) Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining* 4(1): 1–14.

Castells M (2011) *The Rise of the Network Society: The Information Age: Economy, Society, and Culture, Vol. 1*. New York: John Wiley & Sons.

Cicourel AV (1964) *Method and measurement in sociology*. Free Press.

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

Edwards A, Housley W, Williams M, et al. (2013) Digital social research, social media and the sociological imagination: Surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodology* 16(3): 245–260.

Garland D (2001) *The Culture of Control: Crime and Social Order in Contemporary Society*. Oxford: Oxford University Press.

Gayo-Avello D (2012) ''I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper'' – A balanced survey on election prediction using Twitter Data. arXiv preprint arXiv:1204.6441.

Giddens A (2002) *Runaway World: How Globalisation is Reshaping Our Lives*. London: Profile Books.

Ginsberg J, Mohebbi MH, Patel RS, et al. (2009) Detecting influenza epidemics using search engine query data. *Nature* 457(7232): 1012–1014.

Harris M (2012) Oh man, the crowd is getting an F in social science. *Daily Crowdsource*, June. Available at: http://dailycrowdsource.com/crowdsourcing/articles/opinions-discussion/1158-oh-man-the-crowd-is-getting-an-f-in-social-science.

Housley W, Williams M, Williams M, et al. (2013) Special issue: Introduction. *International Journal of Social Research Methodology* 16(3): 173–175.

Howard PN, Duffy A, Freelon D, et al. (2011) Opening closed regimes: What was the role of social media during the Arab Spring? Project on Information Technology and Political Islam. Working paper 2011.1.

Lewis P, Newburn T, Taylor M, et al. (2011) Reading the riots: Investigating England's summer of disorder. The London School of Economics and Political Science.

Lidén G and Nygren KG (2013) Analysing the intersections between technology, performativity, and politics: The case of local citizen dialogue. Transformations 23.

Mills CW (1959) *The Sociological Imagination*. New York, NY: Oxford University Press.

Mislove A, Lehmann S, Ahn YY, et al. (2011) Understanding the demographics of twitter users. *ICWSM* 11: 5.

Pak A and Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: *LREC*, Malta, May.

Procter R, Vis F, Voss A, et al. (2011) Riot rumours: How misinformation spread on Twitter during a time of crisis. *Guardian*. Available at: http://www.guardian.co.uk/uk/interactive/2011/dec/07/london-riots-twitter.

Procter R, Vis F and Voss A (2013a) Reading the riots on Twitter: Methodological innovation for the analysis of big data. *International Journal of Social Research Methodology* 16(3): 197–214.

Procter R, Crump J, Karstedt S, et al. (2013b) Reading the riots: What were the police doing on Twitter? *Policing and Society* 23(4): 413–436.

Procter R, Housley W, Williams M, et al. (2013c) Enabling social media research through citizen social science. In: *ECSCW 2013 Adjunct Proceedings*, 3, Cyprus, September.

Procter R, Housley W, Williams M, et al. (2014) Social media: A new technology for civic participation? In: *International conference on social media and society*, Toronto, September.

Ruppert E, Law J and Savage M (2013) Reassembling social science methods: The challenge of digital devices. *Theory, Culture & Society* 30(4): 22–46.

Savage M (2013) The 'Social Life of Methods': A critical introduction. *Theory, Culture & Society* 30(4): 3–21.

Savage M and Burrows R (2007) The coming crisis of empirical sociology. *Sociology* 41(5): 885–899.

Sayer A (1992) *Method in Social Science: A Realist Approach*. Hove: Psychology Press.

Shearing C and Wood J (2007) *Imagining Security*. London: Willan.

Sloan L, Morgan J, Housley W, et al. (2013) Knowing the tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological Research Online* 18(3): 7.

Stahl B, Eden G and Jirotka M (2012) Responsible research and innovation in information and communication technology. In: Owen R, Bessant J and Heintz M (eds) *Responsible Innovation*. Chichester: Wiley & Sons.

Stepanova E (2011) *The Role of Information Communication Technologies in the ''Arab Spring''. Implications beyond the Region*. Washington, DC: George Washington University (PONARS Eurasia Policy Memo no. 159).

Thelwall M, Buckley K and Paltoglou G (2011) Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology* 62(2): 406–418.

Tinati R, Halford S, Carr L, et al. (2013, in press) The promise of big data: New methods for sociological analysis. *Sociology*.

Tumasjan A, Sprenger TO, Sandner PG, et al. (2010) Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM* 10: 178–185.

Urry J (2003) *Global Complexity*. Cambridge: Polity.

Williams R, Stewart J and Slack R (2005) *Social Learning in Technological Innovation: Experimenting with Information and Communication Technologies*. Cheltenham: Edward Elgar Publishing.

Williams ML, Edwards A, Housley W, et al. (2013) Policing cyber-neighbourhoods: Tension monitoring and social media networks. *Policing & Society* 23(4): 461–481.