THE PROMISE OF BIG DATA: NEW METHODS FOR SOCIOLOGICAL ANALYSIS

Ramine Tinati, Susan Halford, Leslie Carr, Catherine Pope
University of Southampton,
United Kingdom

INTRODUCTION

The current emergence of 'Big Data' is both promising and challenging for social research. As well as offering scale, way beyond most previous social science resources, these data capture social activity in real-time, over time, providing details on what people do and say 'in the wild', rather than what they say they do in interviews and surveys[i]. Meanwhile, the digital nature of these data open new potentials for data mining and linking (boyd and Crawford 2011; Halford, Pope and Weal 2012). However, Big Data also raises some serious challenges for social research, not least methodologically. Indeed, we suggest that – so far – the scope for harnessing Big Data to social science research has been limited by the methods in use.

We can work our argument through the example of Twitter, the micro-blogging website. Launched in 2006, by 2011 Twitter had over 300m users and 200m tweets daily. This success resonates with recent social science interest in 'mobilities' – how the social emerges in dynamic flows of people, objects, images and information (Urry 2000) – and specifically with 'network society' (Castells 1996) in which information – now the key commodity – flows across time and space between loosely connected individuals and groups that form and re-form fluid identities and connections transcending older ties of place, time, class, gender, race, and so on. Networks, in this sense, do not reflect society but rather shape or even produce society (Urry 2000). The social is assembled (Latour 2006) in the everyday practices that constitute the 'global networks' of multinational enterprises and the heterogeneous, uneven and dynamic 'global fluids' '… of people, information, objects, money, images and risks that move chaotically across regions in strikingly faster and unpredictable shapes' (Urry 2000; 190).

However, the potential for Twitter data (or other social media and digital data) to address these questions will demand new methodological approaches. To date, social science researchers have drawn largely on purposive or random samples from Twitter with some small scale content analysis of these. Sampling in this way predefines important actors and/or renders all actors equal as members of a random sample, denying the possibility of tracing which actors and information emerge as important over time. Small scale content analysis allows in-depth analysis but no possibility

of understanding where and how this content or these users are positioned within the broader Twitter stream. Rendering Big Data manageable in this way overrides its nature as 'big' data, by-passing 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.

In this paper we present a new tool for harvesting and analysing Twitter data, underpinned by a broader set of methodological considerations, which begins to address some of these limitations. We work our case through an analysis of the Twitter activity surrounding the recent student fees protests in the UK, we show how the combination of quantitative and qualitative analysis within a broader methodological approach that draws on 'wide data' might help to connect Twitter research more firmly with emblematic sociological concerns with networks, mobilities and flows.

METHODOLOGICAL PRINCIPLES

Our approach has been driven by the following underlying principles. First, *begin with the network*. If we are interested in the on-going flow of information and action, we need tools that can explore how these emerge within the network, rather than imposing a priori assumptions about who or what is important, or using sampling frames from beyond the network to make the data manageable. Second, we must *capture the dynamic flow* of tweets, to explore the network as it grows. Third, we must *overcome methodological polarisation between macro and micro analysis*: between large-scale metrics – which measure the structures and patterns of Big Data – and analysis of micro-level interactions – the communications of individuals (Larsson and Moe 2011), allowing the combination of technical capabilities with in-depth qualitative research methods. From these principles we have developed a computer-based tool that enables us to explore the metrics, dynamics and content of Twitter network formation and information flows, at both macro and micro levels.

#nov9

In what follows we analyse retweets in the #nov9 Twitter stream that draws together tweets around the rise in student fees in the UK in 2011. The total collection contains 12,831 tweets made by 4737 Twitter users 8th October 2011 – 21st November 2011. The time series of these tweets are shown in Figure 1. We consider: what information is flowing? Which actors are most widely cited? How well connected are the tweeters? And do these change over time? Specifically, we analyse the most retweeted tweets (100+) to

explore what information flows and how and examine the roles that emerge in the network over time. In this way we can *trace what the emergent network produces, rather than using the network as a data source to observe actors or tweets selected in advance.* Our analysis is based on the visualisation at http://youtu.be/KvdmdQkS-CM and 'network snapshots' in Figure 2, which represent the retweet communication network which correspond to sections of the analysis below.

In the Flow: information and actors

The red nodes in the video and Figure 2 identify the users who have received a significant number of retweets, whilst the 'edges' (or links from these red nodes) show who has retweeted them, and subsequent retweets. It is immediately obvious that there are only a small number of highly retweeted users. These are not necessarily the most prolific tweeters but their place in the flow of information is clearly significant. The visualisation also shows that a third of the highly retweeted users were apparent a week before the protest, and by 9am on 9th November, 9 of the 12 were already present. In short, the noise of total information flow is often dominated by the voices of a few who, once they have gained a voice, increase their audience and therefore volume over time. As the network of communication grows, it becomes harder to become popular. Whilst several of these users might be characterised as 'the usual suspects' our method reveals less known figures who acquired significance in the network as it grew over time.

Alongside the temporal pattern in user popularity, there was a shift in the content of the highly circulated information over time, from initial calls to participation to later discussion of police tactics. The single most retweeted post attached photographs of the police in action and came from a user with no apparent political affiliation and a relatively small number of followers (c.600) although his chain dissipates within 24 hours. The longest chain, also highlighting policing tactics, lasts 4 days, and was posted by a Guardian journalist with over 8000 followers.

Emergent Network Roles

Turning attention from the information flowing to the retweeters themselves we can see some interesting roles emerging. Specifically there *amplifiers* (blue nodes): users who may tweet very little themselves but pushing information on to new audiences, often very swiftly. Analysis of the #nov9 network reveals one particularly active user in this respect. '@REALsocialnet' was the first to retweet three of the four most highly retweeted messages, initiating the wider circulation of these original posts. However, this amplification role was *selective,* with emphasis on the organization and coordination of the protest. Notably, in each case, the retweeter promoted their *own* activities, thorough links to other hashtags and websites. A

hyperlink – if opened – extends the information circulated via Twitter way beyond the original 140 character tweet, for instance, linking to Facebook pages with extensive information and tens or hundreds of thousands of users. Whilst the action extends the flow of the original tweet it also piggy-backs other interests onto this. As the original tweeters gain dominance in the network, they carry with them the retweeter's information, gaining a wider audience for this too.

A second important role that emerges in the network is that of *'aggregator'* (yellow nodes). This is also a retweet activity, but here the contribution is not in being the first to retweet, but in retweeting posts from diverse streams of information, building bridges between discrete networks, pulling threads of information into a single channel. This works in two ways. First, the aggregators are compiling a selected stream of #nov9 tweets for their followers who are not themselves following #nov9, pushing the information on to a wider audience. Second, the aggregators are doing this across multiple hashtag data streams, operating as a node in the wider Twitter network beyond #nov9.

In sum, the combined effects of these emergent roles network led to a complex interconnected network, dominated by a few highly retweeted individuals, whose position strengthens over time, narrowing down the information in flow, specifically – in this case – to concentrate on concerns about the policing this protest. Our analysis shows that this patterning to the flow of information emerges from multiple iterative actions, not only those of the original tweeters – although these are clearly important – but also by the retweeters and aggregators whose selections come to shape the dominant discourse of the network.

CONCLUSION

Rather than selecting users either purposively or randomly, our method allows us to explore which users and which information rise to the surface in an emergent Twitter network. Our tool enables us to move between the macro-structure of the network to the micro-level of individual users and tweets. Our research shows *for the first time* how specific pieces of information flow and how the incremental actions of individual users produce social roles and networks inside Twitter.

In methodological terms, this is just the beginning. There is more we could do within Twitter, looking at the relationship between tweets, retweets and followers for instance, or developing methods that connect related hashtag streams. Beyond this, our research calls for a 'wide data' approach, making links across digital sources e.g. from Twitter to Facebook or online corporate media – would allow us to explore the relationality of these data.

Furthermore, as others have suggested, we need to move beyond the digital, making links to print media as well as data from interviews and observations to develop fuller understandings.

In broader terms, these methodological developments have disciplinary implications. As the zeitgeist shifts towards 'data driven' research we must ensure that social scientists bring their theoretical and epistemological expertise to bear on the field. Not least, we need to insist that these data are not naturally occurring or unmediated – but are socio-technically constructed, produced and represented through particular methods and artefacts. Furthermore, examining their meaning requires robust methodologies, nuanced conceptual vocabularies and theoretical frameworks. However, we must be frank about the capacity of our existing methodological repertoire, which may not be sufficient in this endeavour. Indeed, as Savage (2010) concludes of more generally, in his analysis of post-war British sociology, the future of the social sciences may depend on building 'intellectual and technical alliances' with other ways of knowing, not least of which – we suggest – in the context of Big Data will include the computational sciences.

REFERENCES

boyd, D., and Crawford, K. (2011) *Six Provocations for Big Data*, presented at the Oxford Internet Institute 'A Decade in Internet Time: symposium on the dynamics of the internet and society', September 21, 2011.

Castells, M. (1996) *The Rise of Network Society: The Information Age* Oxford, Blackwell.

Halford, S., Pope, C., and Weal, M. (2013) *Digital Futures? Sociological challenges and opportunities in the emergent Semantic Web Sociology 47 (1) pp.173-198*

Larsson, A., and Moe, H., (2011) ' Studying political micro-blogging: Twitter users in the 2012 Swedish election campaign' *New Media and Society* 14(5) pp.729-747.

Latour, B. (2006) *Reassembling the Social* Oxford, OUP.

Savage, M. (2010) *Identities and Social Change in Britain since 1940: the politics of method* Oxford, Oxford University Press.

Urry, J. (2000) 'Mobile sociology' *British Journal of Sociology*, 51(1), 185–203.

Endnotes

[1] This is not to suggest that big data is somehow 'pure' or 'free' of social norms and constraints simply that these data are produced beyond rather than through sociological research methods.

FIGURES

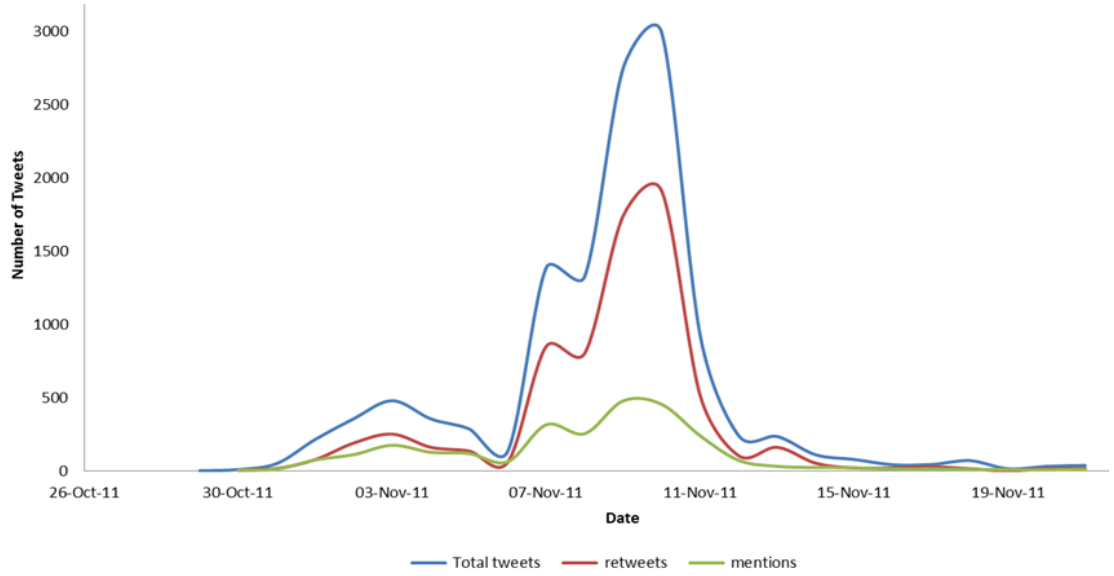
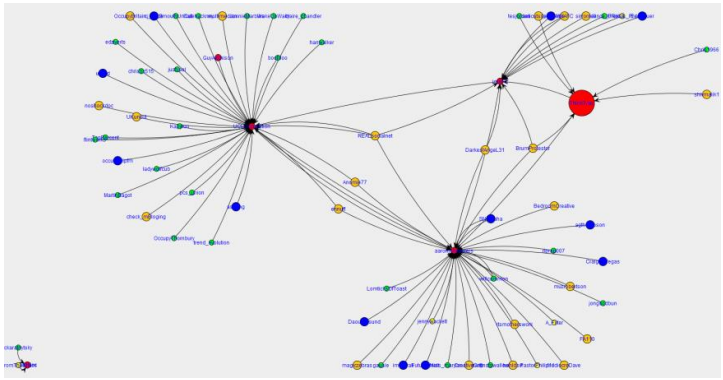
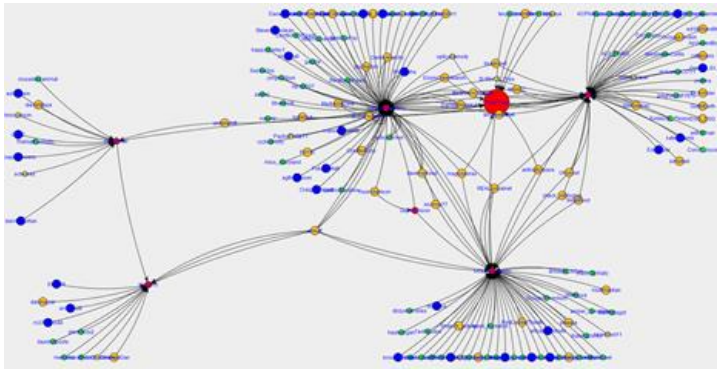


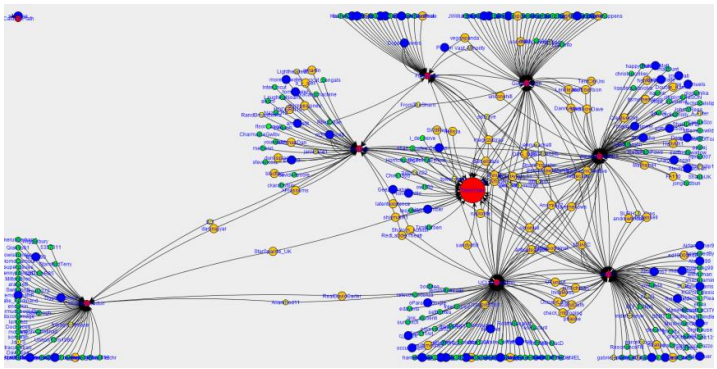
Figure 1: Number of #nov9 Tweets, Retweets and Mentions 30th October - 19th November 2011



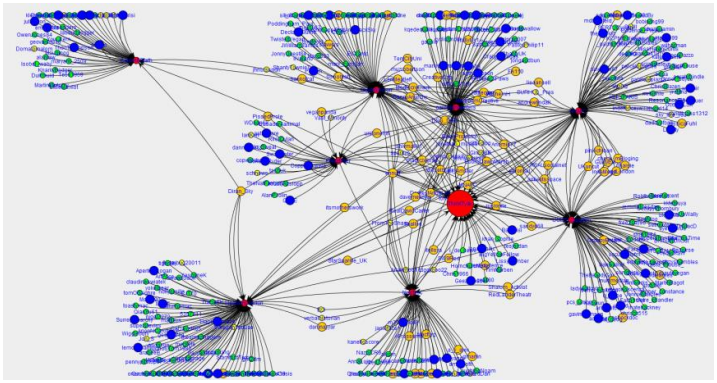
21:00 Nov 3rd 2011



21:00 Nov 7rd 2011



21:00 Nov 8rd 2011



09:00 Nov 9rd 2011

Figure 2: #nov9 Retweet Network