Big Data: Methodological Challenges and Approaches for Sociological Analysis

Journal:	*Sociology*
Manuscript ID:	SOC-Mar-2013-ARTC-045
Manuscript Type:	Article
Keywords:	big data, twitter, methodology, networks, information flow

Big Data: Methodological Challenges and Approaches for Sociological Analysis

The current emergence of big data is both promising and challenging for social research. This paper suggests that realising this promise has hitherto been restricted by the methods that have been applied in social science research, which undermine our potential to apprehend the qualities that make big data so appealing not least in relation to the sociology of networks and flows. With specific reference to the micro-blogging website Twitter, the paper outlines a set of methodological principles for approaching these data that stand in contrast to previous research and introduces a new tool for harvesting and analysing Twitter built on these principles. We work our argument through an analysis of Twitter data linked to political protest over the rise in UK University fees. Our approach transcends earlier methodological limitations to offer original insights into the flow of information and the actors and networks that emerge in this flow.

Key Words: *Big Data, Twitter, methodology, networks, information flow*

Introduction

The current emergence of 'Big Data' is both promising and challenging for social research. Originally coined to describe digital data sets so large that they required non-standard computational facilities and software for storage and analysis (Manovich 2011) as data generation has grown, as has the capacity of standard computers, the term now encompasses a wider range of remarkable properties inherent in these data. Beyond the scale of these data *per se* attention is drawn to their proportionality – these are 'whole' data sets, capturing everything within a particular field (e.g. utility records) or on a particular platform (e.g. Twitter) (Hale and Margetts 2012); they are dynamic – capturing social activity in real time, over time; and they offer information on what people do and say 'in the wild', rather than what they say they do in interviews and surveys[i]. The digital nature of these data also opens up new potentials for data mining and data linking, allowing connections to be made between diverse data (boyd and Crawford 2011; author 2012)

However, Big Data also raises some challenges for social research. These are emergent, but it is clear that there are new and important ethical issues to deal with (Neuhaus and Webmoor 2011). Furthermore, in between the enthusiasm of some – Latour (2007) suggests '… it is as if the inner workings of private worlds have been pried open' (p.2) – and the scepticism of others, for whom these data are ephemeral froth distracting us from more serious sociological endeavours, lie some important ontological and epistemological questions: what do these data represent and what claims can be made from them? Finally, before we can address either of these issues fully, there are methodological challenges. Indeed, until we know how to apprehend and analyse Big Data, we cannot appreciate the range or scale of ethical and epistemological questions that may arise; and will arise variously across different forms of Big Data. For although the term may imply coherence and uniformity, 'Big Data' is not one thing but many, differentiated *inter alia* by content, structure, ownership and availability. While much has been made of the potential of Big Data for social research (e.g. Savage and Burrows 2006 and subsequent debate), for reasons of privacy and/or commercial sensitivity, many of these datasets remain in the hands of governments and private corporations.

One significant exception to this is Twitter, the micro-blogging website whose content is (almost entirely[ii]) public, visible to anyone who chooses to search and follow users, and available via Twitter's own Application Programming Interface (API), which - depending on the methods used – allows access to a (1) small selection of the tweets via the search or streaming service, (2) the '*garden-hose*', a 10% random sample, or (3) the '*firehose*' all tweets made. Not surprisingly, Twitter has generated a considerable amount of interest amongst social scientists: since its launch in 2008, there have been over 110 scholarly publications about Twitter (International Bibliography of Social Sciences Accessed 08/10/12). Whilst little of this has been published in mainstream sociology (see Murthy (2012) in this journal), there is much here to interest sociologists, for instance in attention to practices of impression management, micro-celebrity and personal branding (Jackson and Lilleker 2011; Marwick and boyd 2011; Hargittai and Litt 2011); and to questions of participatory democracy and political mobilization (Grant et al 2011; Larsson and Moe 2011; Tufekci and Wilson 2011; Segerberg and Bennett 2011; Saunders et al forthcoming).

However, 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 *viz.* 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* of tweets. 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 what follows, we elaborate our claim that Twitter research remains limited by its methodological approaches. Specifically, we suggest not only that social scientific approaches have failed to capture the most interesting qualities of Big Data but also that, because of this, we cannot make the most of these data to address currently emblematic sociological concerns. In this paper we present a new tool for harvesting and analysing Twitter data, underpinned by a broader set of methodological considerations, which together begin to address some of these limitations. Working 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 sociological analysis.

Sociological and Methodological Challenges

Twitter was established in 2006 as a micro-blogging website, allowing individuals to 'tweet' 140 character messages made immediately visible in the timelines of their 'followers' and to anyone searching the Twitter website. By 2011 Twitter had over 300m users and 200m daily tweets (http://blog.twitter.com/2011/06/200-million-tweets-per-day.html). The emergence and success of Twitter resonates with some of the recent cutting-edge concerns of sociology. At the meta-level, it is symptomatic of a wider transition away from the 'social as society' – at least, as society bounded by nation states – towards the 'social as mobility', emergent in dynamic flows of people, objects, images and information (Urry 2000). More specifically, this can be characterised as a '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). For Urry (2000) these fluids have no clear point of departure or arrival, no necessary end-state and are characterised by '… emergent, unintended and non-linear consequences' (ibid; 195) *that sociology, as a discipline, should find better ways of interrogating.*

Big Data in general, and Twitter as a specific example, offer hitherto unexplored potential for empirical work of this type. Twitter holds promise not least because of its availability to researchers but also its openness. It is easy to access and the conversations between users are relatively easy to follow as are the users' networks of 'friends'. Unlike other social networking websites such as Facebook, Twitter does not enforce reciprocal relations between users, enabling registered users to 'follow' as many others as they wish, whether or not they 'follow back'. Bill Gates, for instance, follows 140 Twitter users, but has nine million followers. Anyone can observe any post on Twitter, even without being an identified 'follower'; messages can be sent to any other user using their unique @username (known as 'mentions') – and these are displayed publicly on all profiles and tweets; and users can pass on any tweet – via the 'retweet' function. Users may also choose to group their discussions around particular topics or events using hashtags (e.g. #election2012) within the body of the tweet[iii]. In short, Twitter has no defined structures beyond the technical functions of tweet (140 characters), 'follow', @mentions, and retweet. In principle then, we can follow the emergent flow of information – what is tweeted, retweeted and hashtagged, and the evolving networks that form and reform between people over time.

Twitter has already been the subject of some fascinating research, particularly in political science, but the methods used mean this work is somewhat limited in scope and, understandably, this research has not engaged with the issues raised by the sociology of networks, mobilities and flows. For instance, research on the role of Twitter in grass-roots activism and its potential for enhancing participatory democracy, *either* pre-selects the important actors (elected politicians especially) *and/or* takes a sample of tweets or users within a defined area of activity, often a hashtag stream. Of course, if the aim of the research is to explore how elected politicians use Twitter then pre-selecting these actors is an entirely consistent choice. Furthermore, it is inevitable that the quantity of Twitter data will require some management, and since hashtags emerge from user practices this makes them a sampling frame. However, if the aim is to explore which actors are active and influential on Twitter during election debates, or what kinds of networks emerge between actors at this time, we need to take *the network itself* at the starting point. Sampling tweets, whether purposively or randomly, denies the opportunity to trace which actors and information emerge as important over time. Rather, this method predefines which actors are important and/or renders all actors equal as members of a random sample. Nothing can be said about what the network itself produces.

Similarly, small scale content analysis of selected tweets or studies of particular users (for instance, 30 tweets from each of 60 Twitter accounts (Waters and Williams 2011) or following the Twitter stream of 51 MPs (Jackson and Lilleker 2011) allows in-depth analysis but no possibility of understanding where and how this content or these users are positioned within the broader Twitter stream. More generally, previous research has neglected the dynamic nature of information flows and network connections on Twitter. In one exception, the number of tweets around a hashtag is reported at 13 weekly intervals (Segerberg and Bennett 2011) but the wider temporality of the network itself – who is connected with whom via direct messages or retweets – is not reported. Notably however, in explaining Twitter temporality, Segerberg and Bennett (2011) make links to contemporaneous events off-line, highlighting the value of making links across data sources, flagging up the importance of following hyperlinks within tweets, whilst others (Hargittai and Litt 2011; Marwick and boyd 2011) have begun to use mixed methods to evaluate Twitter usage. These are important methodological developments to which we return in a moment. For now the point we want to make is this: *whilst the key characteristics of Big Data are its scale, proportionality, dynamism and relationality, the methods that have been used in social science to date have fallen some way short of enabling us to explore this.*

Meanwhile, Twitter has also attracted attention from computer scientists. Compared with the 110 articles on IBSS there are over 350 articles with a primary focus on Twitter listed in the Association for Computing Machinery (ACM) Digital Library. In contrast to social science research, computational approaches endeavour to capture data at scale, through the development of algorithms and technical solutions (Cohen et al 2009) that aim to reduce the computational times of data processing (Dean and Ghemawat 2004) or improve mechanisms for aggregation and storage to enable faster and more efficient access (Herodotou et al 2011). However, interest from the computer sciences has not been confined to technical concerns. As Manovich (2011) suggests, the advent of Big Data makes easier for those with the appropriate programming skills and knowledge of various social media APIs, to ask social questions. Indeed, there is a stream of such research on Twitter from computer science exploring, for instance, friendship networks (Macskassy and Michelson 2011), political orientations (Conover et al 2011) and the diffusion of information (Chang 2010; Bakshy et al 2012). This might support Savage and Burrows' (2006) claim that the availability of new forms of data is moving the centre of gravity for social research away from sociology, although it is important to note that attention is more often to observing patterns and network structures *per se* rather than exploring meaning or explanation. Where claims to social knowledge are made these take the form of 'big' claims about the patterns in Twitter, for example using Natural Language Programming and sentiment analysis to search for key words to determine the 'happiness' of a tweet (Dodd 2011), or an individual's political affiliations (Rao 2010). These types of approaches favour computational techniques rather than theoretically informed or conceptually nuanced sociological analysis, let alone fine grained qualitative analysis.

Nonetheless, computational research brings relevant methods to the study of Twitter. In particular, the capacity to apprehend Big Data and analyse network structures, to measure the volume of data and the flow of information and relations between actors. These techniques are not untried in sociology. Indeed, from the application of graph theory to social ties (Moreno 1953) that gained particular momentum from the quantitative 'revolution' of the 1960s (Barnes 1969) to John Scott's pioneering work to embed Social Network Analysis (SNA) in the sociological methods repertoire (1998; 2000) these techniques have a long history in the discipline. However, whilst some of these techniques have been used by political scientists to research Twitter (Larsson, 2011) they remain untried in its sociological analysis. Further, echoing our critique above, the established SNA techniques have some limitations. As Scott (2008) has argued, the power of SNA would be improved if it were to move beyond static metrics and statistical measures of network structures and connectivity, to expose the temporal nature of the data and this, we suggest, is particularly pertinent here. In what follows we present a new software tool, developed to meet these challenges.

The Method

Our tool provides a dynamic visualisation of the information flows and social networks that emerge in Twitter over time. Its development was driven by the following underlying principles. First, *start with the network*. If we are interested in the actors and outcomes that are produced in the on-going flow of information, 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. Our second principle is that 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 (see also Larsson and Moe 2011; and Edwards 2010), allowing the combination of technical capabilities with in-depth qualitative research methods.

From these principles we have developed a computer-based tool that enables the metrics, dynamics and content of Twitter information flows and network formation to be explored in real-time or via historic data. This uses some common techniques and metrics from SNA, for example measuring static properties such as number of nodes (users), edges (directed communications between one user and another), in-degree (the number of directed tweets or retweets towards an individual user) and out-degree (measuring the mentions made or retweets by that user of another user). Beyond this, our tool also enables us to (i) examine the dynamic properties of Twitter networks, incorporating an adaptable graphical user interface to visualise this; (ii) develop associated metrics to measure the flow of information at scale and over time; and (iii) 'zoom in' to examine the content of conversations and communications between individuals.

Remaining true to these principles does *not* mean that we have to engage with the whole Twittersphere. Although some have done this (Ahn, 2007), the scale and heterogeneity makes this difficult. Rather, our tool filters the data stream following the primary principle: that is, starting with the network itself, drawing on user generated hashtags. Hashtags are produced to link a tweet to a particular topic, effectively a 'bottom-up' curation of tweets around a particular topic into a single stream of data. Second, the tool uses an algorithmic filtering solution to reduce the volume of data based on the characteristics that individuals display within the network: the number of times they have tweeted, the number of times they have retweeted or been retweeted, their connectivity within the network and the role that they play in the diffusion of information. It is important to focus on the retweet function because it is the means by which information is diffused across Twitter. User 'A' tweets: this is seen by their followers and anyone else who searches Twitter for that user or topic, or happens to come across the tweet serendipitously. If User 'B' retweets the original post, then this is seen by all User B's followers (etc.), who may in turn retweet. And so on. Following retweets allows us to trace the *flow of information*, rather than simply observe individuals or tweets, which we have no way of knowing whether anyone has read, let alone passed on to anyone else. The retweet also offers a way to observe which information and which actors become important as the network evolves: *what the network produces, rather than using the network as a data source to observe actors or tweets selected in advance.* In what follows we demonstrate our method through an analysis of the #nov9 Twitter network that draws together tweets around the rise student fees and the day of protest that took place on November 9^{th} 2011.

Political Activism on Twitter

There has been a great deal of interest in if and how Twitter might be used to facilitate political activism and, perhaps, engender new forms of grass-roots mobilization and enhance participatory democracy. Some dramatic claims have been made in public debate, not least about the 'Obama paradigm' of electioneering via social media (Theocharis 2011), reference to 'Twitter revolutions' in the Arab Spring (Sullivan 2009) and the role of Twitter in the 2011 urban riots in the UK (http://www.huffingtonpost.co.uk/2011/08/08/london-riots-twitter-that_n_920791.html). However, there is relatively little concrete or detailed evidence about the actual role of Twitter in these events (Theocharis 2011), leading to calls for systematic research beyond 'anecdotal evidence and sweeping generality' (Segerberg and Bennett 2011; 199). We need to know more about how Twitter is used in practice and take care to avoid the abstraction fallacy (Segerberg and Bennett 2011) that ascribes political features to a technology, rather than exploring these with rigorous investigation. Rising to this challenge, Theocharis (2011) and Segerberg and Bennett (2011) provide theoretically informed and empirically detailed accounts of Twitter use in political protest showing that Twitter is used both to mobilise and inform over sustained periods as well as more tactically during actual demonstrations (Theocharis 2011); and that Twitter plays an important role in connecting diverse networks of people, although this is done in different ways in different hashtag streams and changes over time (Segerberg and Bennett 2011). Both studies conclude that Twitter is expanding the portfolio of strategies available for organization, becoming one tool in an increasingly sophisticated political communication and organization repertoire.

This emphasis on the place of Twitter in the broader political ecology is important, but previous research remains methodologically limited because of its focus on a small number of tweeters, pre-defined on the basis of institutional affiliation (Theocharis 2011) or random samples of a larger set of data (Segerberg and Bennett 2011). Thus, whilst these studies might begin with the data generated by a hashtag they impose their own external criteria on this to sample data, rather than tracing the actors and relationships that emerge in the network. Second, the analysis is based on static properties of the network and the user – for example, the 'friends' and 'follower' metrics for key actors – rather than the analysis of the relations that emerge within the network over time. Even though Segerberg and Bennett take counts at several points in time, we cannot see the dynamics that produce these statistics. Finally, whilst there is some inclusion of qualitative data – illustrative tweets for example – the methods employed cannot allow us to see what information (which tweets) are moving across the network or linking users together. To redress these concerns, in what follows we present our approach which allows us to explore which users, information and linkages emerge within the network over time.

#nov9

Our dataset is the set of tweets using the hashtag #nov9 linked to political protest against the rise in university tuition fees in England and in particular the demonstration that took place in London on November 9th 2011. The total collection of tweets, harvested via Twitter's streaming API service, contains 12,831 tweets made by 4737 Twitter users 8th October 2011 - 21st November 2011. These data identify the author, time of tweet and the content of the tweet. Figure 1 provides the basic metrics from this data stream, showing an uneven flow of activity over time, with a flash of activity around the day of the protest that then tails-off[iv].

Figure 1 about here

Over 54% of the tweets are retweets – passing on others' messages – whilst only 18% of all tweets contained 'mention' messages to another user, showing a high re-circulation of information intended for a general, rather than specific, audience.

From these metrics a series of questions emerge. What information is flowing? Which actors are most widely cited? How well connected are the tweeters? And how do these change over time? In our analysis we focus primarily on the retweet network because this allows us to trace the information in flow, the actors involved and the networks that emerge as a consequence, although – as will become apparent – this also engages us with direct messaging between users and with other sources of data, beyond the tweet itself. In the sections below we use our tool to filter the data archived from #nov9 to trace tweets that have been retweeted 100+ times[v]. This is a *data-driven* simplification of our data that allows us to trace the most widely flowing streams of information. Figure 2 shows a series of static snapshots taken from November 3rd – November 9th, visualising the growth of the retweet network over this time whilst the *conversation playback video* at http://youtu.be/KvdmdQkS-CM shows a dynamic visualisation of the network over the same period that can be paused at any point in time. In what follows, we explore the flow of information across this retweet network and examine the roles that emerge in the network over the time.

In the Flow: information and actors

The red nodes in Figure 2 and the *conversation playback video* identify the users who have received a significant number of retweets within the data being examined, whilst the 'edges' (or links from these red nodes) show who has been retweeting them, and any subsequent retweets of this message.

Figure 2 about here

It is immediately obvious that there are only a small number of highly retweeted users (in these data, only 0.26% or 12 individuals were retweeted more than 100 times). These are not necessarily the most prolific tweeters – their average tweet-retweet ratio is 1:12 – so they would not have been identified on these grounds alone, but their place in the flow of information is clearly significant. Four of these users were already apparent almost a week before the protest, and by 9am on 9th November, 9 of the 12 were already present, showing the emergence of consistent key players who, indeed, only consolidate their role as central nodes in the network over the period. In contrast to previous research that identifies the interesting actors as a way of sampling data, our method means that these key players are derived from the network itself. Significantly, whilst several of these users might be characterised as 'the usual suspects' there are also some less known figures.

'@UCLOccpation' for instance, linked to students at University College London and highly retweeted in our network also features at the top of Theocharis' (2011) chart of Twitter accounts associated with University student politics. Similar to this, '@imAFC' (now 'NCAFC_UK') represents the 'National Campaign against Fees and Cuts', a coalition of students and workers who actively protest against student tuition fees apparently linked to the antticuts.com website. On the other hand, '@ThinkTyler', 'and '@aaronjohnpeters' are individuals, one linked to a grassroots group working towards improved democratic engagement and the other with no apparent organized affiliations. These tweeters became important in the network, but would not have been captured by sampling well-known actors and, given the number of tweeters in the network, might not be selected by random sampling methods.

As the network grows, and new highly retweeted users emerge, this heterogeneity narrows. By November 7th the most highly retweeted messages are from known anti-cuts tweeters and these users maintain their ranking throughout the remainder of our analysis. This phenomenon can be described by the concept of preferential attachment (Barabasi and Albert 1999). 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. Prior to the protest on November 9th, 4 individuals were identified as highly retweeted, and this number quickly rose to double that within 5 days. However subsequently, the rate of growth of individuals to become highly retweeted decreased, and instead, the already highly retweeted individuals reinforced their voice within the network, although they were not necessarily adding new tweets. As Figure 2 illustrates, 24 hours after the protest, the nodes with the highest amount of edges (here, retweets) become more popular and gain more edges. At this stage, the flow of information within the network becomes saturated with the tweets of these highly retweeted individuals, overshadowing the unknown users and their tweets.

Alongside this temporal pattern in user popularity, we see a temporal turn in the content of the highly circulated information around #nov9, from calls to participation to discussion of police tactics and apparent evidence of police brutality.

> *[Wed 02 Nov 2011 20:40:49] "RT @aaronjohnpeters: There is a march of 10000 students to the city of London on November 9th come! #frontline #Nov9"*

> *[Tue 01 Nov 2011 12:36:38] "RT @UCLOccupation: @pcs_union will you and your members join and support the November 9th demonstration by students http://t.co/LTvspyra #nov9"*

> *[Sat 05 Nov 2011 20:27:52] "RT @Fitwatcher: More disgusting police behaviour. We need to think seriously about #nov9 and how we're going to defend ourselves. #acab"*

> *[Mon 07 Nov 2011 16:55:40] "RT @HeardinLondon: In case you missed the news The Met have given the go ahead to use bullets on the kids if they misbehave too much at the student demo #Nov9"*

Of the Top Ten most retweeted posts, nine concerned policing and allegations of brutality. Figure 3 shows the retweet chains for these top 10 posts. The single most retweeted post, from @ThinkTyler – a user with no apparent political affiliation and a relatively small number of followers (c.600) – begins *'I got warned not to post these pictures …'* suggesting an appetite for using Twitter as a mechanism of direct defiance, although the chain dies away within 24 hours. In comparison, the longest chain, also highlighting policing tactics sustains itself over 4 days, and was posted by a Guardian journalist with over 8000 followers.

Figure 3 about here

Attention to the number of followers of an individual (re)tweeter is important, since any post will show up in the timelines of all those followers. The more followers, the more widely the information is circulated. This point is compounded if we consider the URLs embedded in many of the tweets above. A hyperlink – if opened - extends the information circulated via Twitter way beyond the original 140 character tweet. For instance, the URL embedded in the UCLO tweet above links to a Facebook page which itself contains over 31,000 users. Making use of this 'wide data' underscores points made elsewhere about the importance of placing Twitter within a broader ecology of tactics available for political mobilization (Segerberg and Bennett 2011) and enables us to place specific political mobilizations on a broader canvas.

Emergent Network Roles

In our analysis so far, we have concentrated on the users that emerged as highly retweeted as the network grew, and the content of their messages: the information that flowed. It is not, however, the actions of these original tweeters which cause the information to flow. In this section, we turn our attention to the role of the retweeters: the users who pass on information and who may come to occupy a particularly significant role in the emergent network. Figure 2 and the conversation playback video http://youtu.be/KvdmdQkS-CM show that the pattern of retweets is not random or evenly spread across the network. Specifically, there are users who – whilst not particularly active in generating content themselves – play an important role in passing information on, being the first to retweet, pushing information on to new audiences, often very swiftly (these are identified as the blue nodes in Figure 2 and the online visualisation, Figure 3 shows the speed of retweeting). Retweeting is not spread evenly across users but, rather the flow of information is strongly shaped by these '*amplifiers'*. Analysis of the #nov9 network reveals one particularly active user in this respect. Throughout the lifetime of the network '@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:

> [Sun 30 Oct 2011 16:47:01] *"RT @UCLOccupation: http://t.co/7D8jFsRE debut is in UCL's Jeremy Bentham room - 9 pm Wednesday #realsocialnet #nov9 #solidarity"*

> [Tue 01 Nov 2011 16:39:36] *"RT @imAFC: London regional meeting TONIGHT at 6pm in UCL. Also remember @REALsocialnet also at UCL from 8pm. #realsocialnetwork #nov9"*

> [Thu 03 Nov 2011 12:11:14] *"RT @UCLOccupation: National Student demonstration to the City of London November 9th http://t.co/Ggm4PadM #ukuncut #nov9 #weareeverywhere"*

In every case, the retweeter promoted their *own* activities, thorough links to other hashtags and websites. 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',* identified by the yellow nodes in Figure 2. 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. Some individuals such as '@REALsocialnet' take on both the role of the *'amplifier'* and *'aggregator'*, for example '@ennuff' who is a first retweeter and and aggregates posts from a number of highly retweeted individuals, assimilating potentially valuable information.

> [Sat 05 Nov 2011 21:50:41] *"RT @aaronjohnpeters: New November 9th Student Demo (and more) website launched! http://t.co/j4tdFmpy Please RT! #nov9 #siteworker #occupylsx"*

> [Wed 09 Nov 2011 22:05:03] *"RT @ThinkTyler: I got warned not to post these pictures "all over the internet"; as it will "infringe their job". http://t.co/OAXLx944 #9nov #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.

Conclusions

The primary aim of this paper has been to consider the methodological challenges that face those interested in engaging with Big Data, specifically – in this case – with Twitter data, and to demonstrate a new research tool designed to address some of these challenges. We have worked our case through an analysis of the #nov9 data stream and our findings extend previous research in the following ways. Rather than selecting users either purposively or randomly, we examine the emergence of a communications network, and have explored which users and which information rise to the surface as a result of the dynamic flow of information. To achieve this, our method enables us to 'zoom' from analysis of the macro-structure of the network – where our analysis is based on quantitative algorithmic methods – to the micro-level of individual users and tweets. This allows us to see how information diffuses and flows between users over time, and to explore the networks that emerge as a consequence. Whilst previous research concentrates on content and aims to link this to off-line activities 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.

This shows, very clearly, that Twitter is not one thing but many. Twitter is neither a medium for news nor a method of organizing but both: its form is contingent produced in the multiple iterations of users. In the spirit of Science and Technology Studies we might say that Twitter is in a process of becoming. It is not a defined set of possibilities, however tightly specified the technical platform may be. Even within a particular hashtag, the information stream and the users involved are heterogeneous and dynamic. Original tweeters cannot know the fate of their posts, whether they will capture a wider imagination, or be selected by the influential retweeters and aggregators. Similarly, the retweeters and aggregators can aim to promote particular information but once they have done this the fate of their retweets will be in the hands (or thumbs) of others. Nonetheless, dominant discourses may emerge within a hashtag stream, even if they dissipate and disappear just as quickly as they appear. Although the #nov9 hashtag was used extensively previous and subsequent to the demonstration, overall this was only a short period of time; the networks themselves are dynamic, fluid and changing (Urry 2000), the topics that the #nov9 hashtag represented at the time of the protest no longer resonates through the Twitter network, instead – and is bound to change yet again – it now represents no subject or use, just random tweets from disconnected individuals.

In methodological terms, this is just the beginning. Whilst our tool offers some important advantages over previous methods, there is clearly more that we can do. Within Twitter analysis, we might explore the relationship between retweets and followers to trace flows of information and emergent linkages between users. That is, are retweets only made by those who are *already* followers of a particular user, or by others who come across the tweet either thorough the hashtag or in other more serendipitous ways? We might also aim to develop methods that connect hashtag streams, rather than following one stream only. Indeed, there are still question that need to be asked with regards to who actually sees this information, and retweets only offer one way to explore the exposure of information within a communications network. Beyond Twitter, we have seen the importance of joining digital traces, but the methods we have for doing this are still in their infancy. We also need to consider other types of Big Data. Whilst we have focussed on Twitter, the principles that we propose - tracing the flow of information, following links within the data, and theoretically supporting the data – will be important in other contexts. Other sources of Big Data, whether containing human interactions or machine transactions, need to be apprehended in their full state in order to benefit from the wealth of information they contain.

Pushing methodological development in this direction has enabled us to engage with (one example of) Big Data, paying attention to its scale, proportionality and dynamism whilst our emphasis on the importance of 'wide data' – links across digital sources e.g. from Twitter to Facebook or online corporate media – begins to open up the relationality of these data. Furthermore, as others have suggested, we should make links beyond the digital to print media, interviews and observations and so on. These methodological developments have both epistemological and ethical implications. Our capacity to engage with whole data sets at scale, and to combine qualitative with quantitative analysis may go some way to allying concerns about the status of Twitter data, by allowing us to position individual tweets/tweeters, or samples, within the whole network; by allowing us to follow the flow of data – where it goes – rather than simply comment on the existence of these data; and by allowing us to engage with detailed content as well as overall patterns. This is 'wide data' rather than Big Data and we suggest that this methodological approach will also serve to strengthen our claims to knowledge. Meanwhile, however, a wide data approach raises some new ethical questions. Building links across data sets can pose profound threats to individual privacy (Bizer 2009). Whilst individuals posting on Twitter are likely to be aware that this will be publically available, they may not consider the composite picture that can be built up about them by combining multiple sources. This is not something that we have done in this paper, but it will become increasingly possible and presents challenges to us as social scientists as to how we will govern our practice in ethical ways.

REFERENCES

Ahn, Y., Han, S., Kwak, H., Moon, S., & Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. *Proceedings of the 16th International ACM World Wide Web Conference* (pp. 835–844).

Alterman, J., (2011) 'The revolution will not be tweeted' *The Washington Quarterly* 34(4), pp.103-16.

Barnes, J., (1969) 'Graph theory and social networks: a technical comment on connectedness and connectivity' *Sociology* 3(2), pp. 215-32.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. International *Journal on Semantic Web and Information Systems* 5(3), pp.1-22.

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.

Dodds, P. S., Harris, K. D., Isabel, M., Bliss, C. A., & Danforth, C. M. (2011). 'Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter' *LoS ONE* 6(12): e26752.

Grant, W., Moon, B., and Grant, J., (2010) 'Digital Dialogue? Australian Politicians' use of the Social Network Tool Twitter' *Australian Journal of Political Science* 45(4), 579-604.

Hale, S. A., & Margetts, H. (2011). Understanding the Mechanics of Online Collective Action Using "Big Data". *SSRN Electronic Journal*, 1–7.

Hargittai, E., and Litt, E., (2011) 'The tweet smell of celebrity success: explaining variation in Twitter adoption among a diverse group of young adults' *New Media and Society* 13(5) pp. 824-842.

Jackson, N., & Lilleker, D., 'Microblogging, constituency service and impression management: UK MPs and their use of Twitter' *Journal of Legislative Studies*, 17(1), pp. 86-105.

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.

Latour, B. (2007) 'Beware, your imagination leaves digital traces' *Times Higher Literary Supplement*, (April).

Moreno, J., (1953) *Who Shall Survive? Foundations of sociometry, group psychotherapy and sociodrama* New York, Random House.

Marwick, A., and boyd, D., (2011) '"I tweet honestly, I tweet passionately": Twitter users, context collapse and the imagined audience' *New Media and Society*, 13(1) pp. 114-133.

Manovich, L. (2011) 'Trending: The Promises and the Challenges of Big Social Data' *Debates in the Digital Humanities*, 1–17.

Murthy, D. (2012) 'Towards a sociological understanding of social media: theorizing Twitter' *Sociology* 46(6) 1059-1073.

Neuhaus, F., and Webmoor, T., (2012) 'Agile ethics for massified research and visualisation' *Information, Communication and Society* 15(1), pp. 43-65.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying Latent User Attributes in Twitter. *Proceedings of the 2nd international workshop on Search and mining user-generated*. New York, New York, USA: ACM Press.

Savage, M., and Burrows, R., (2007) 'The coming crisis of empirical sociology' *Sociology*, 41(5) pp.885-899.

Segerberg, A., and Bennett, L., (2011) 'Social media and the organization of collective action: using Twitter to explore the ecologies of two climate protests' *The Communication Review* 14(3), pp. 197-215.

Scott, J. (1998) 'Social Network Analysis' *Sociology* 22 (1), pp. 109-127.

Scott, J. (2000). *Social Network Analysis: A Handbook*. London, Sage.

Scott, J. (2008) 'Network Analysis' in Darity, W. (Ed) *International Encyclopaedia of the Social Sciences* Gale Virtual Reference Library. Gale.

Sullivan, A. (2009) 'The revolution will be Twittered' *The Atlantic* 13 June 2009.

Theocharis, Y (2012) 'Cuts, tweets, solidarity and mobilisation: how the internet shaped the student occupations' *Parliamentary Affairs* 65, pp. 162-94.

Tufekci, Z., & Wilson, C. (2012) 'Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square' *Journal of Communication*, 62(2), 363–379.

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

Ward, J. (2011) 'Researching citizens online' *Information, Communication and Society* 14(6) pp.917-36.

Waters, R., and Williams, J. (2011) 'Squawking, tweeting, cooing, and hooting: analyzing the communication patterns of government agencies' *Journal of Public Affairs*, 11(4), pp. 353-63.

Endnotes

[i] 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.

[ii] It is possible to 'protect' tweets from other users unless they are identified followers. We are aware of no information on how often this is done, but it appears to be very rarely indeed.

[iii] Aalthough notably the # function evolved from bottom-up use rather than being an original Twitter function.

[iv] It is worth noting that there are no constraints on the use of a particular hashtag for any specified purpose. #nov9 has been used more recently to refer to a range of events and topics, not only the student fees protest. However, since our tool allows us to view the content of the tweets using this hashtag over the archived time period and to see which information 'rises to the top' in terms of number of retweets we can be confident that the vast majority of the data collected refers specifically to the student protests.

[v] Our tool allows us to set this filter at any level. We have chosen 100+ here to allow us to see the detail in this particular set of data and address the questions that we are focussing on in this paper. For other data or other questions the level might well be set lower, or indeed higher.

Figures

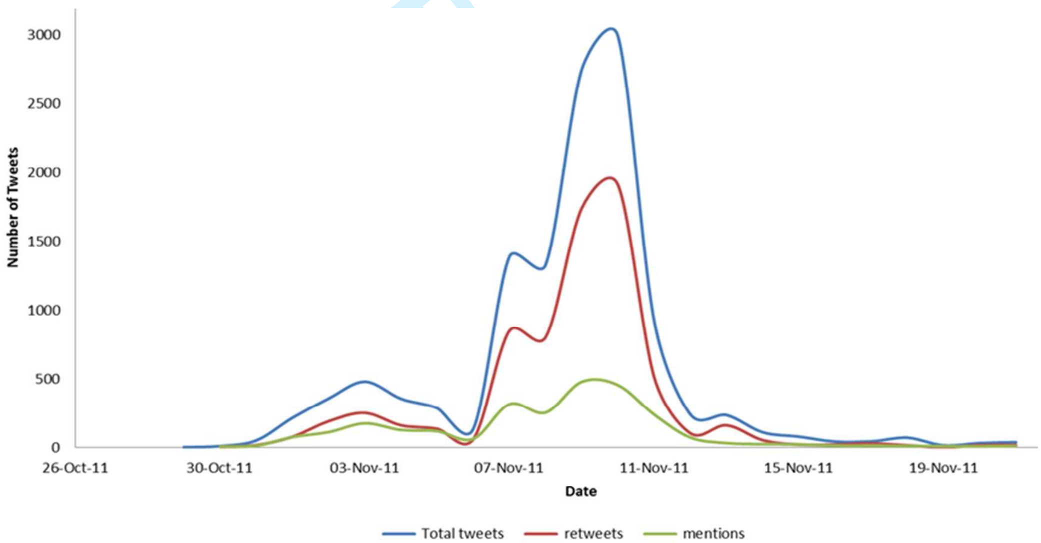
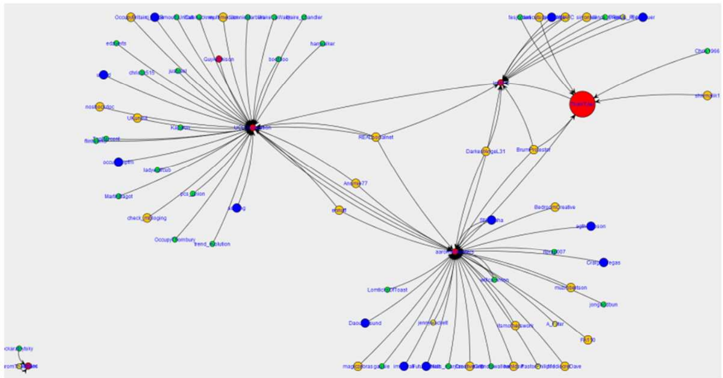
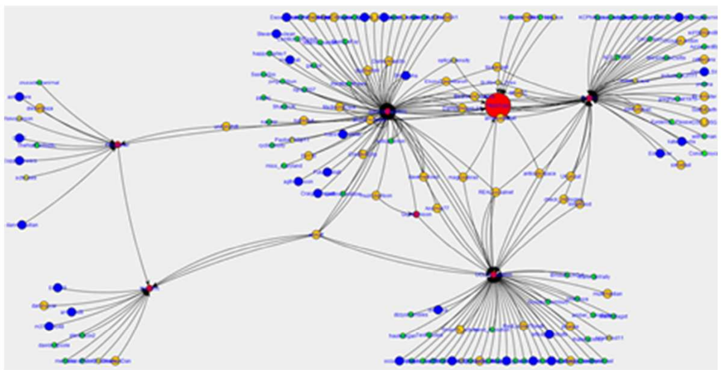


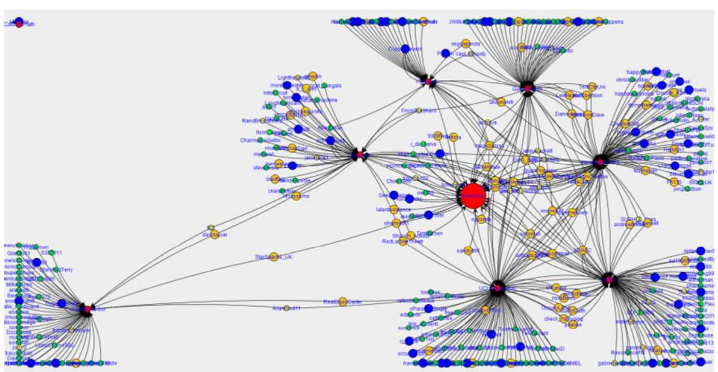
Figure 1: Number of #nov9 Tweets, Retweets and Mentions
30th October - 19th November 2011



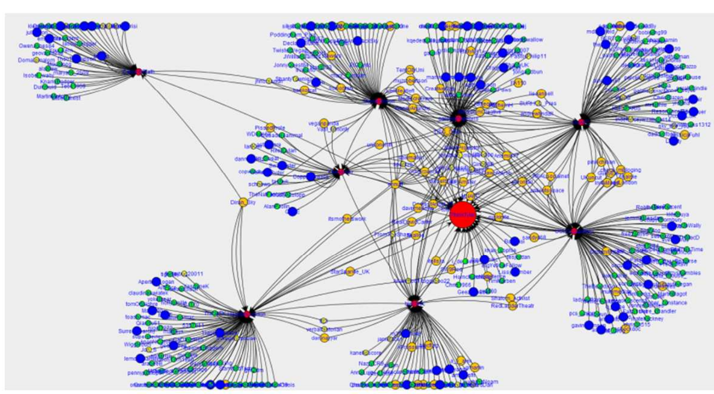
21:00 Nov 3rd 2011



21:00 Nov 7rd 2011



21:00 Nov 8rd 2011



09:00 Nov 9rd 2011

Figure 2: #nov9 Retweet Network

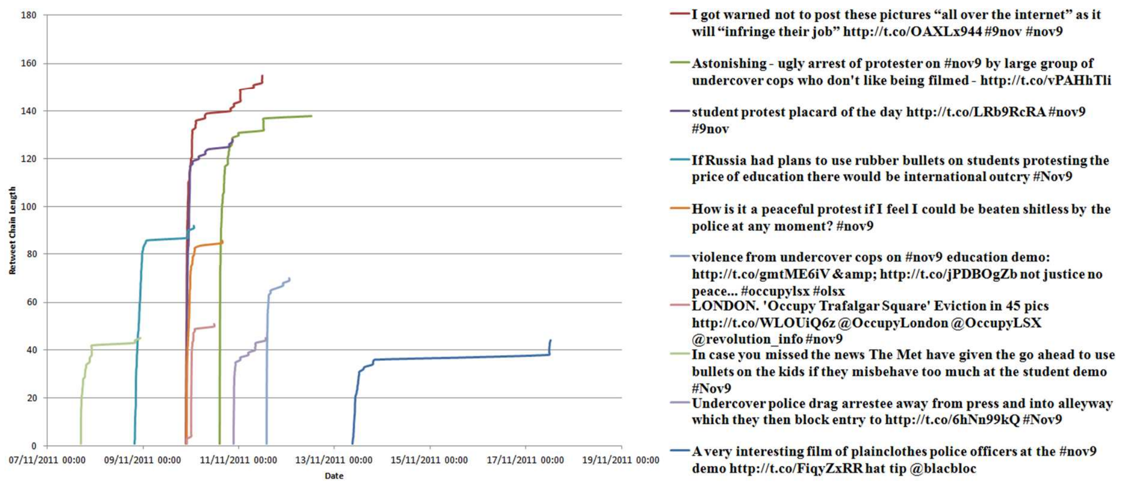


Figure 3: #nov9 Ten Longest Retweet Chains