Interrogating Big Data for Social Scientific Research: an Analytical Platform for Visualising Twitter

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

rt506@ecs.soton.ac.uk

The recent emergence of 'big data' – large scale digitized data sets capturing commercial and transactional activity – is both promising and challenging for social scientific research. On the one hand, these data offer information on 'action in the wild' – the things that people actually say and do, rather than what they say they do in surveys or interviews – and they do so at a scale rarely, if ever, approached by conventional social scientific research methods (Savage and Burrows 2007). On the other hand, these data pose a range of methodological, theoretical and philosophical challenges. How can we access and describe these data, make them manageable for social scientific research? What do these data show? And what are the ethics of working with these data?

In this paper, our way into these debates is methodological: we begin with a practical approach to the harvesting and visualization of big data, one way to render big data analysable for social scientific research. Our empirical focus is on Twitter, the micro-blogging website, which allows individuals to 'tweet' 140 character messages which are made immediately visible in the timelines of their 'followers' and are also searchable by any other Twitter account holder. Established in 2006, Twitter has grown at a phenomenal rate with more than 300 million users and 200 million tweets daily[1] and has evolved new functionalities to allow users to tweet directly to each other, using @username within tweets, and to group tweets around shared topics, using #hashtags that can be set up by anyone, on any subject. In sociological terms, Twitter is a dynamic social network offering the immediately visible trace of apparently spontaneous social interactions and relationships.

Research on Twitter to date has concentrated heavily on content especially the role that Twitter communication has played in political and social movements (Ward 2011) often emphasising the links between on-line and off-line activities and the relevance of Twitter in place (despite its cyberspace platform) although this is a contested point in Twitter research (Takhteyev, Gruzd, and Wellman 2011). Although there has also been some attention to the nature of the ties between social media users, including Twitter users, (Haythornthwaite and De Laat 2010) so far relatively little attention has been paid to the nature of Twitter networks themselves: how they emerge and take shape or the pattern of interaction between users. Whilst more traditional social network analysis has paid considerable attention to these questions off-line the data used and analysis, to date, have been static (Scott 2010), paying little attention to the dynamics of social networks.

To begin to answer these issues our paper presents a new software tool, developed to provide visualisation of the social networks that emerge in Twitter hashtag discussions over time. Twitter provides an excellent case study for the development of new computational tools which take advantage of its well-structured API, which allows a large amount of complete data to be collected in a fairly short amount of time and; unlike other social networking sites, it places only limited restrictions on rate of data collection. Our data collection involved a harvest of tweets made using the #nov9 hashtag over the period from 21th October 2011 to 21st November, during this period, 12831 tweets and their supporting metadata were collected and stored in chronological order, and also included information about author and creation time of the tweet, the tweet text, and also additional identifier information. The #nov9 hashtag represents the national British political protest against the rise in student university tuition fees on November 9th 2011. This was chosen in particular for the burgeoning literature on political activism on Twitter, helping providing a compelling and relevant case study to demonstrate our methodological approach to harvesting and visualising big data in order to conduct social scientific research.

Using the data collected, our software enables a timeline of communications around the specific hashtag to be visualised, based on retweets – when an original tweet is re-sent by another user – and named mentions of users – when one user cites the @username of another in a tweet. Analysing these uses of Twitter allow us to explore the influential tweets and tweeters within a given network, and to examine how a network forms over time. Our data will be presented using dynamic film of the hashtag communications as they unfolded over time. This provides a visual analytical approach to understanding the flow of information within a network, and additionally enabling the communications to be 'paused' and zoomed into, examining at certain points in time the actions of the actors, at not only the individual level, but their role within the network as well. This is not only providing a detailed understanding of the communications between actors, but exploring the pathways and flow of information pushes the methodological approach to understanding big data in terms of its dynamic nature, something that helps explain the translation and processes of a network (Scott, 2010).

Beyond this, our methodological approach to big data is also exposing the potential power structures that exist within an online network of individuals communicating, revealing their potential reflections with those in the non-virtual world. Exposing the underlying communication network structures reveal certain properties help distinguish and separate individuals from each other, not only by looking at their position within the network, but also their relevance in terms of the dynamic network activities.

[1] http://blog.twitter.com/2011/06/200-million-tweets-per-day.html

Finally, the drive towards big data and social scientific research raises a number of ethical questions regarding the conventional guidelines towards using private and public information and the issues surrounding informed consent. Reflecting on the ethical challenges that Neuhaus *et* al. (2011) discuss, we examine the future implications that our methodological approach for big data may expose.

1. ACKNOWLEDGEMENT

This research was funded by the Research Councils UK Digital Economy Programme, Web Science Doctoral Training Centre, EP/G036926/1.

2. REFERENCES

Haythornthwaite, Caroline, and Maarten De Laat. 2010. “Social Networks and Learning Networks: Using social network perspectives to understand social learning.” Pp. 183-190 in *7th International Conference on Networked Learning 2010*, edited by Lone Dirckinck-Holmfeld et al. University of Aalborg Retrieved (http://www.lancs.ac.uk/fss/organisations/netlc/past/nlc2010/abstracts/PDFs/Haythornwaite.pdf).

Neuhaus, Fabian, Timothy Webmoor, and Park End Street. 2011. “Information , Communication & Society Agile Ethics For Massified Research And Visualization.” *Society* (March 2012):37-41.

Savage, M, and R Burrows. 2007. “The Coming Crisis of Empirical Sociology.” *Sociology The Journal Of The British Sociological Association* 41(5):885-899. Retrieved (http://dx.doi.org/10.1177/0038038507080443).

Scott, John. 2010. “Social network analysis: developments, advances, and prospects.” *Social Network Analysis and Mining* 1(1):21-26. Retrieved March 7, 2012 (http://www.springerlink.com/index/10.1007/s13278-010-0012-6).

Takhteyev, Yuri, Anatoliy Gruzd, and Barry Wellman. 2011. “Geography of Twitter networks.” *Social Networks* 34(1):1-25. Retrieved (http://linkinghub.elsevier.com/retrieve/pii/S0378873311000359).

Ward, Janelle. 2011. “Reaching Citizens Online.” *Information Communication Society* 14(6):917-936. Retrieved (http://dx.doi.org/10.1080/1369118X.2011.572982).