Interrogating Big Data for Social Scientific Research - An Analytical Platform for Visualising Twitter

Ramine Tinati, Susan Halford – rt506@ecs.soton.ac.uk – Web and Internet Science – University of Southampton

(1) Research Background and Introduction

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. The World Wide Web, an evolving and fast changing socio-technical network of activity offers a vast resource of big data, especially since the change towards Web 2.0 and the Social Web.

Big data is offering researchers from a variety of disciplines a resource which was once difficult or simply impossible to obtain, yet, it must be approached with some level of caution, as a number of challenges are also attached to its capabilities (boyd and Crawford 2011). How can we access and describe these data, and make it manageable for social scientific research? What skills are needed and technologies available to be able to effectively and efficiently use this new kind of data? And most importantly, what do the data show?

In response to these challenges, this paper presents a new software tool, developed to provide a dynamic visualisation of the social networks that emerge in Twitter hashtag conversations 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.

To demonstrate the capabilities of the tool, we analyse a corpus of Tweets using the #nov9 hashtag, which represents the UK political protest against the rise in student university tuition fees on November 9th 2011.

(2) Visualising and Filtering Twitter Networks

We have developed a computer-based tool which enables the dynamics of conversations between users within a Twitter conversation network to be explored in real-time or via historic data. The tool provides an observational window to explore the interactions and communications as they occur. This draws upon some of the common techniques and metrics used within social network analysis, including the analysis of the communication network in terms of its static properties: number of nodes, edges, network degree, etc.

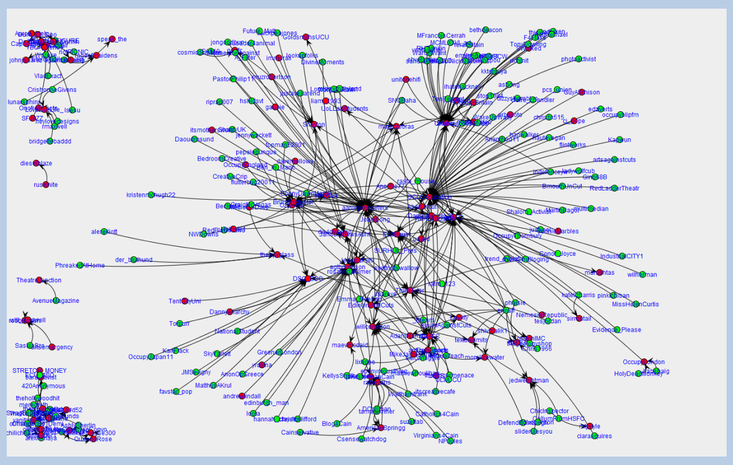
The tool also provides a way into exploring the dynamic properties and evolution of the network. It has been designed to process both live, real-time streaming data from Twitter or harvested datasets containing Twitter messages, typically associated with a specific hashtag or conversation topic.

Finally, to overcome the problems with visualising and analyzing such large amounts of information, a algorithmic solution[20] has been developed to enable the communication network to be filtered based upon a selection of characteristics that individuals exhibit within the network, including the number of times they have tweeted, retweeted, their connectivity within the network, and the role they play in the diffusion of information.

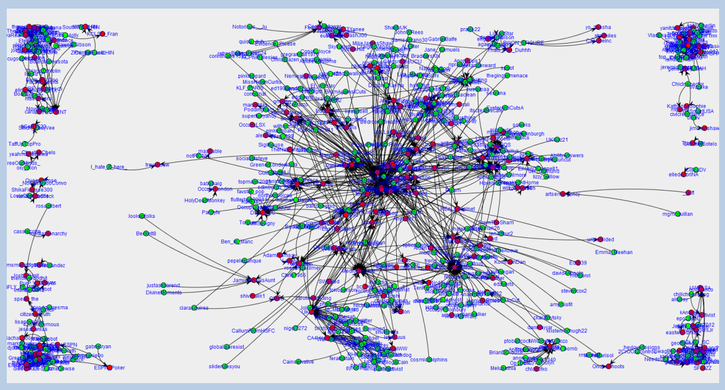
Using this approach, it becomes possible to work with a large set of communications, both visually and analytically; users, which represent the nodes within the network – are displayed differently – in their size and color – which not only offers a simplified visualisation of the network, but also provides a method to identify specific communication patterns between different clusters of individuals; which is further illustrated via the dynamic display of communications.

(3) - Exploring the November 9th Protests – Filtered vs. Unfiltered Retweet Conversation Stream

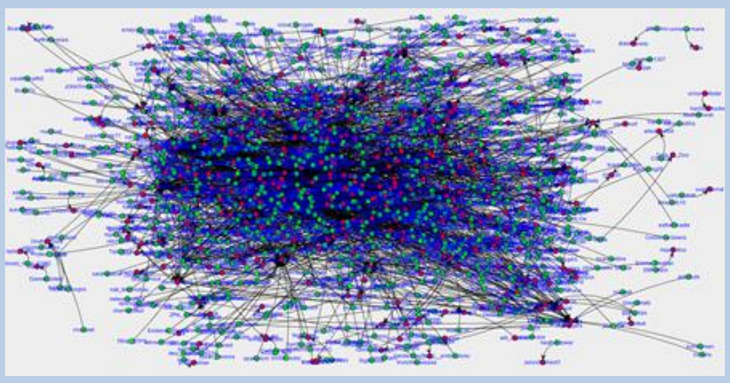
Fig 1. Unfiltered Retweet Network



(a) Retweet Network as of Nov 3rd 2011

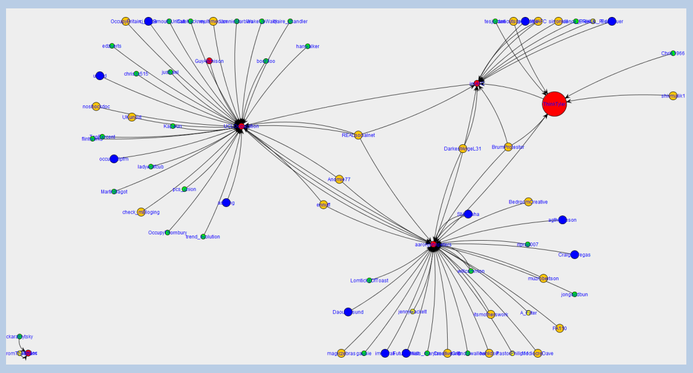


(b) Retweet Network as of Nov 6rd 2011

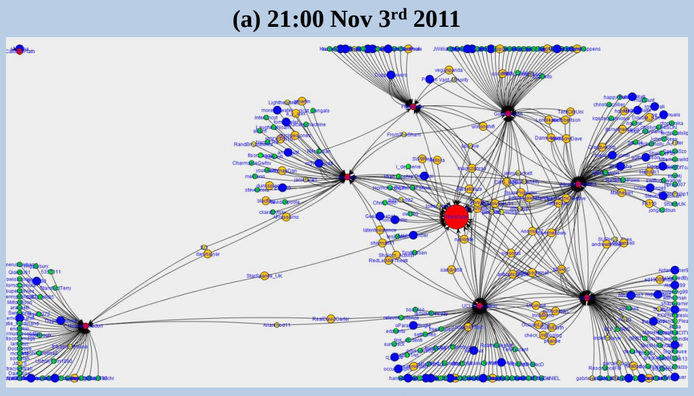


(c) Retweet Network as of Nov 9rd 2011

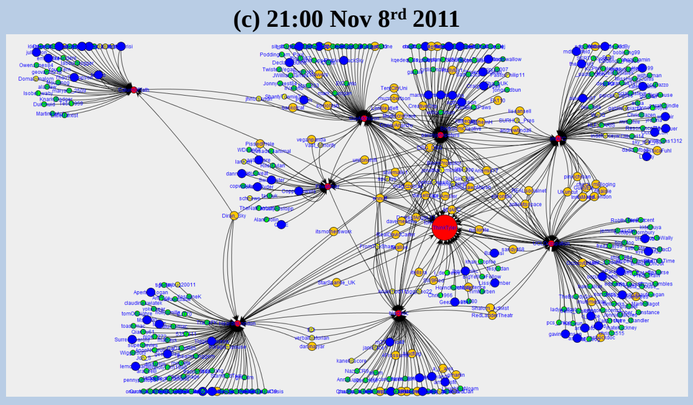
Fig 2. Filtered Retweet Network



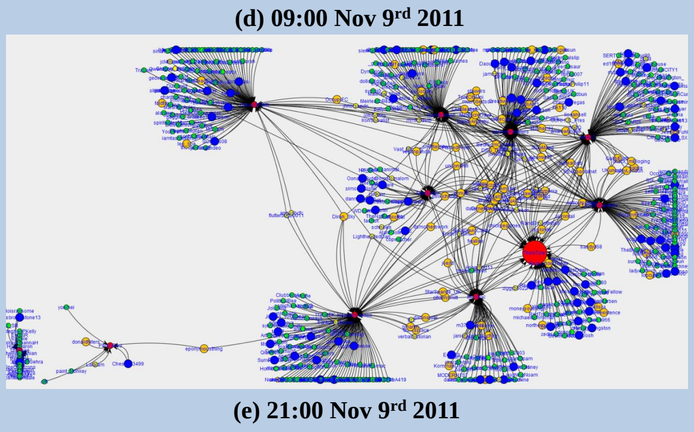
(a) 21:00 Nov 3rd 2011



(c) 21:00 Nov 8rd 2011



(d) 09:00 Nov 9rd 2011



(e) 21:00 Nov 9rd 2011

Figure 1 represents the retweet network as a 'flat' network, in its original, unfiltered state. As Figure 1a-c – which are visualisation of the network of individuals (the nodes) and the retweets (the edges) – illustrates, examining the dynamic growth of the network when the network in an unfiltered state provides very little observational benefits, let alone analytical use.

As the network time slices shown in (a) to (c) show, even during the early stages of the networks visualisation, observations prove difficult and the rapid growth of the communications make this more difficult as time increases. The clusters that were identifiable before the protest on the 3rd November become impossible to distinguish by the 9th November.

In order to handle the scale of the data, the tool's filtering algorithm offers a way to reduce the complexity of the network, concentrating on the communications between individuals which exhibit certain network characteristics.

Figure 2.a-e represents the growth of the filtered retweet network, which offers a much clearer view of the structure and overall growth of the network due to the applied filtering mechanism. First observations of the network shows that were a number of users whose presence was constant throughout the growth of the network. These users, which are represented by the red nodes, are the individuals who had received over 100 retweets, and their size is determined by the number of retweets that they have received in comparison to the other red nodes