How Long Is a Tweet?
Mapping Dynamic Conversation Networks on *Twitter* using Gawk and Gephi

Assoc. Prof. Axel Bruns
ARC Centre of Excellence for Creative Industries and Innovation
Queensland University of Technology
Brisbane, Australia
a.bruns@qut.edu.au — @snurb_dot_info — http://mappingonlinepublics.net/

Abstract

Twitter is now well-established as the world's second most important social media platform, after *Facebook*. Its 140-character updates are designed for brief messaging, and its network structures are kept relatively flat and simple: messages from users are either public and visible to all (even to unregistered visitors using the Twitter Website), or private and visible only to approved 'followers' of the sender; there are no more complex definitions of degrees of connection (family, friends, friends of friends) as they are available in other social networks.

Over time, *Twitter* users have developed simple but effective mechanisms for working around these limitations: '#hashtags', which enable the manual or automatic collation of all tweets containing the same #hashtag, as well allowing users to subscribe to content feeds that contain only those tweets which feature specific #hashtags; and '@replies', which allow senders to direct public messages even to users whom they do not already follow.

This paper documents a methodology for extracting public *Twitter* activity data around specific #hashtags, and for processing these data in order to analyse and visualise the @reply networks existing between participating users – both overall, as a static network, and over time, to highlight the dynamic structure of @reply conversations. Such visualisations enable us to highlight the shifting roles played by individual participants, as well as the response of the overall #hashtag community to new stimuli – such as the entry of new participants or the availability of new information. Over longer timeframes, it is also possible to identify different phases in the overall discussion, or the formation of distinct clusters of preferentially interacting participants.

Introduction

Twitter is now well-established as the world's second most important social media platform, after *Facebook*. It differs from *Facebook* in a number of important aspects, of course: its 140-character updates are designed for brief messaging, and its network structures are kept relatively flat and simple: messages from users are either public and visible to all (even to unregistered visitors using the *Twitter* Website), or private and visible only to approved 'followers' of the sender; there are no more complex definitions of degrees of connection (family, friends, friends of friends) as they are available in other social networks.

By contrast, the 'globally public by default' nature of tweets lends itself to the development of means for automatically organising discussions of specific topics through shared conversation markers. Over time, *Twitter* users have developed the simple but effective convention of using '#hashtags': shared keywords or abbreviations preceded by the hash symbol '#' which enable the manual or automatic collation of all tweets containing the same #hashtag, as well allowing users to subscribe to content feeds that contain only those tweets which feature specific #hashtags. (Technically, it should be noted, #hashtags are no more than a specially formatted form of keyword, and it would be just as easily possible to collate and track all tweets containing a given keyword without the '#' symbol; however, the inclusion of that symbol distinguishes tweets

whose authors deliberately included a specific #hashtag – for example, '#Australia' – from those which merely happen to use the keyword, without hash, in conversation – i.e. 'Australia'.)

Twitter users have also developed a similarly simple mechanism for addressing their public tweets specifically at particular users: here, the name of the recipient is prefixed with the '@' symbol. As this does not rely on dedicated software support, but merely requires knowledge of the addressee's *Twitter* username, it allows senders to send such @replies even to users whom they do not already follow (or even to @reply to non-existent usernames, in order to make a point in conversation), and again demonstrates the flat and relatively barrier-free structure of the *Twitter* network. Notably, as with #hashtags, the @reply convention was introduced as an *ad hoc* measure by the *Twitter* user community first, and then only later supported with enhanced functionality both by the *Twitter* Website and by third-party social networking applications (Halavais & Martin-Elmer, 2009). (The *Twitter* site itself, for example, now enables users to see in one place all the @reply messages they have recently received, and converts #hashtags and @replies in displayed tweets to links to pages with all messages in that #hashtag or @reply conversation.)

By now a ubiquitous practice on *Twitter*, @replies (not least also in combination with #hashtags) lend themselves very obviously to research projects that seek to analyse the network structure of the *Twitter* community or its various subsets. Studies of @reply patterns *within* all tweets marked with a specific #hashtag may help to identify the most central users within that topical network – in doing so also exploring what actual activity metrics may indicate 'centrality' –, while studies *beyond* such specific #hashtag communities are able to address questions over how isolated or interconnected individual #hashtag groups are, as well as approaching the development of a more comprehensive map of global *Twitter* @reply networks that may also highlight broader patterns of clustering along regional, linguistic, demographic, topical, or other lines. Finally, an incorporation of tweeting timeline data into such studies can also trace the ebb and flow of @replying activity over time.

In light of these possibilities, it is somewhat surprising that there appears to be a relative dearth of *Twitter* network studies at present. There is, it should be acknowledged, a growing body of scholarly work which studies *Twitter* as such, from a variety of perspectives; importantly for our present purposes, for example, Honeycutt & Herring (2009) as well as boyd *et al*. (2010) examine the processes of *Twitter* conversations using @replies and retweets. Honeycutt & Herring examine the 'conversationality' of everyday @reply exchanges on *Twitter* and visualise the cascading series of tweet-and-response interactions between multiple participants in @reply discussions, finding "that the responsiveness of Twitter as a conversational environment is at least at the low end of moderate and probably higher" (2009: 6); it should be noted that in contradistinction to the study presented here, these conversations took place without the use of #hashtags, which may be expected to help further coordinate multi-user conversations by providing an additional marker of thematic coherence. Building on this work, boyd *et al*. examine the role of retweets as an element of such conversations, finding that "the practice contributes to a conversational ecology in which conversations are composed of a public interplay of voices that give rise to an emotional sense of shared conversational context" (2010: 1).

The present study adopts this notion of conversationality in *Twitter* messaging, but focusses its analysis specifically on #hashtag conversations: that is, on @replying and retweeting activities which include a given #hashtag in their exchanges. By contrast, follow-on @replies and retweets which do not themselves include the #hashtag are not included in this analysis. The description of such exchanges as 'conversational' is disputable, of course: while any @reply or retweet clearly is a response to a previous message from another user, so that the exchange between the two participants meets a minimal definition of 'conversation', such interactions may remain one-off exchanges rather than constituting more engaged, extended conversations in a fuller sense. From this perspective, the true degree of 'conversationality' of such *Twitter* exchanges may be overstated. At the same time, as noted above, a focus only on #hashtagged exchanges will also miss potential follow-on tweets between users if they do not continue to include the #hashtag in these messages; as a result, we may also expect that the #hashtag focus systematically underestimates the level of conversational exchanges between users participating in the #hashtag community. On balance, therefore, the characterisation of #hashtagged @replies and retweets as 'conversational' can be upheld – with the

qualification that what we analyse here will often capture the beginning more than the conclusion of conversations between two or more *Twitter* users.

This approach also differs from the methodology pursued by Mendoza *et al*. (2010) in their examination of tweeting patterns surrounding the 2010 earthquake in Chile. While studies of #hashtag communities similarly introduce a shared thematic focus (the more or less clearly defined topic to which the #hashtag refers), Mendoza *et al*.'s study "used a filter-based heuristic approach": "we selected all tweets using the Santiago timezone, plus tweets which included a set of keywords … which characterized the event" (2010: 2). Such an approach may be able to capture a greater range of conversations than is possible from a #hashtag dataset alone, provided that geographical, keyword, and other filters are appropriately chosen by the researchers – but such intuitive choices necessarily also introduce significant potential for dispute. Arguably, especially in cases where researchers are unable to cover all the permutations of possible keywords, or where keywords themselves should be expected to change over the course of the communicative event, approaches which do not predetermine a range of relevant keywords, but instead follow the #hashtags set by participating *Twitter* users themselves, may well be preferable.

These and similar *Twitter* studies, then, provide a conceptual framework for understanding conversational uses of *Twitter*: they highlight the role of @replies and retweets as practical mechanisms for, as well as visual markers of, public conversation on *Twitter* (*private* conversations, whether, conducted via direct messaging or through accounts whose tweets are not publicly visible, are not included here, of course), and they provide the foundations for an understanding of the diachronic processes of conversation both in general – especially Honeycutt & Herring (2009) – and in the context of specific crises and other communicative events – as in Mendoza *et al*. (2010). Notably, however, they do not significantly focus on examining the *network* structures which may be identified for such conversational exchanges. Mendoza *et al*. do provide a basic investigation of the propagation of breaking news about the Chilean earthquake (via retweets) through the *Twitter* network (2010: 6), and map follower/followee relationships between the twenty most active users in their dataset, but do not comprehensively examine the conversational networks between all users discussing the event.

Indeed, compared to the number of qualitative and quantitative studies of *Twitter* use which examine other aspects of the site, *network* studies of *Twitter* (whether for the entire platform, or for selected subsets) remain comparatively underdeveloped. The most cited *Twitter* network studies at present (Huberman *et al*., 2008; Java *et al.*, 2007) still date from the early years of *Twitter*, and focus largely on studying the structures of *Twitter*'s network of followers and followees (that is, how users subscribe to one another's activity streams) rather than on @reply networks (how users engage with one another in everyday practice). Given the rapid development of *Twitter*, perhaps the comparatively much slower speed of academic publishing means that the vast majority of up-to-date conversational network studies of *Twitter* have yet to see the light of day in published form.

Additionally, continuing technological changes may also have stunted the development of *Twitter* network studies. The rapid growth of *Twitter* itself necessitates some changes to the methods used by these seminal studies. Earlier, while the *Twitter* userbase still measured in the hundreds of thousands, it may still have been possible to generate comprehensive data on the follower/followee structures across the entire community, or to capture all users' tweets; today, with some two hundred million *Twitter* accounts (Shiels, 2011), it would be difficult even to establish a representative sample covering only a small percentage of all users. *Twitter*'s own attitude to making its data available to researchers and other users further complicates matters: while the platform provides an Application Programming Interface (API) for reliable and straightforward access to structured data on users, tweets, and other activities, and has long supported the growth of an ecosystem of third-party applications (for both end users and researchers) around this API, in early 2011 it began to interpret its API access rules in a significantly more stringent fashion, referring potential users who needed access to very large datasets on *Twitter* users and their tweets to a new commercial access provider, Gnip (Melanson, 2011). As a result of this change in policy, many research projects built in good faith on *Twitter*'s previous relatively permissive policy of granting access to large datasets on a case-by-case basis now find themselves cut off from their data source and unable to afford Gnip's (substantial) commercial access fees.

Certain somewhat smaller-scale approaches to researching *Twitter* network structures – which do not require access to the data licenced by Gnip, but continue to be able to work directly with the *Twitter* API, even in its now more restricted form – do remain possible and continue to generate valuable and important insights, however. The purpose of this paper, then, is to document a methodology for extracting public *Twitter* activity data around specific #hashtags, and to process these data in order to analyse and visualise the @reply networks existing between participating users – both overall, as a static network, and over time, to highlight the dynamic structure of @reply conversations as they unfold over time. This work constitutes early outcomes from a three-year ARC Discovery project researching the processes of public communication using social media in Australia, and will use Australian datasets to demonstrate its processes; its methodological approaches, however, are transferrable to the study of *Twitter* communities in a wide variety of other contexts as well.

(It should also be noted in this context that any research which depends on accessing data from social media or other online platforms through an Application Programming Interface is necessarily dependent on the reliability and performance of that API. Research using the *Twitter* API, as we discuss it here, relies on the API's ability to deliver all relevant tweets in time and without fail – but such reliability is far from guaranteed. Further, there is no opportunity for the researcher to independently verify the performance of the API – that is, to test whether or what percentage of relevant tweets fail to be delivered by the API, – since in the absence of any other comparable access points the API is the *only* mechanism for researchers to access these data at scale and over longer periods of time. The API, in other words, acts as an unavoidable 'black box' between researcher and data source; this complicates the research process and prevents researchers from achieving total certainty about their results, but – short of gaining access to the data through other mechanisms – constitutes an inevitable fact of life.)

Why Map *Twitter* Networks?

Social network analysis and visualisation is now well-established as an interdisciplinary field drawing *inter alia* on contributions from mathematics, computer science, social science, media, communication, and cultural studies, and design; its theoretical frameworks and methodologies have been applied to the study of social connections and interactions across offline and online contexts ranging from family ties through scholarly bibliometrics to social media, to name but a few. Online social networks, in particular, have proven especially interesting to researchers utilising these methods, since the analysis of online social structures can provide insights into the interleaving of human interactions with the technological platforms used to enable and support them: the specific affordances of different social networking sites are also reflected in the structure of the social networks which form around them, if not always in the ways intended by their designers.

Additionally, online social networks appear to exert a special fascination for researchers because they are, for the most part, already rich in readily accessible and apparently objective data: it is considerably easier for the researcher to establish who said what to whom, and under what circumstances, in a large and lengthy public discussion on *Twitter* than it is to generate a comparably comprehensive and accurate dataset for a similar offline interaction. Such apparent simplicity can also be deceptive, however, if online social and communicative interactions are simplistically positioned to represent 'social interaction' as such, without also considering how the specific affordances and limitations of the mediating technologies affect the style, form, and format of communication which is possible in each case. Further, it is important to reject any generalisations from the specific userbase of particular online social media platforms, with its specific demographic structure, to the wider population as such.

Even recognising such *caveats*, however, what the application of social network analysis to the study of interactions in online social networks can provide are detailed, site-specific, insights into the processes of communication between the users of these networking sites. An early practitioner of online network analysis, Richard Rogers (2009; 2010) has long argued that what is necessary here is an approach which puts an "emphasis on natively digital methods" (2009: 5), rather than importing and 'digitising' existing, offline methods. This, he describes as *following the medium*, wherever it may lead: an approach which, rather than merely translating to the online environment some of the long-standing questions of media, communications,

and social science research as they have existed in the past, also allows for the organic emergence of new areas of investigation from the research process itself. As Rogers puts it, "the overall purpose of following the medium is to reorient Internet research to consider the Internet as a source of data, method and technique" (2009: 13). In particular, Rogers notes,

> the Internet may be rethought as a source of data about society and culture. Collecting it and analyzing it for social and cultural research requires not only a new outlook about the Internet, but method, too, to ground the findings. Grounding claims in the online is a major shift in the purpose of Internet research, in the sense that one is not so much researching the Internet, and its users, as studying culture and society with the Internet. (2009: 29)

The picture of a research methodology which emerges from this discussion is two-fold, then: first, a relatively open-ended, exploratory engagement with online objects, developing the 'natively digital' methods which are appropriate to their study and examining – not least through experimental trial and error – what useful and reliable data may be gathered about them and their users; this is the 'follow the medium' stage of the research. Second, the development of new research questions, and new methods of analysis in pursuit of these research questions, which make use of these available data; amongst Rogers's challenges to researchers here is the question "how may one rethink user studies with data (routinely) collected by software?" (2009: 7).

If we take social media and social networking platforms to be appropriate online objects for such research, and assume that through their various Application Programming Interfaces (APIs) we may gain access to rich data about user interactions in these spaces, several major research questions emerge in response to this challenge: first, what is the interrelation between the underlying network structures (networks of 'friendship' in *Facebook*, follower/followee relationships on *Twitter*) and the actual patterns of communication which may be observed? Are there strong tendencies, for example, to communicate only with well-established close connections, or – to the extent that specific platforms allow this – do exchanges with comparative strangers also occur (and how often)? Second, what communicative patterns can be observed – how frequent and how extensive are individual interchanges between users? Do they take place mainly between pairs of users, or also between multiple parties? Third, what content is exchanged? What language is used, and to what extent are non-textual materials (images, audio, video, links to external resources) also exchanged between users? Fourth, how do such communicative practices interconnect with wider contextual factors? How do online social network users track, follow, or respond to news and current events, for example, and do what extent are their communicative exchanges influenced by coverage in other (especially mainstream) media? Additionally, this list of research questions (which is by no means exhaustive) may be applied at various levels of scale and specificity, from examining the interactions of a small, select subset of known users to analysing the actions of a large group of participants, or even of the entire userbase of a particular social networking platform; similarly, studies may focus on brief communicative events or attempt a longer-term tracking of social media participation.

Especially in the latter cases, as researchers are dealing with potentially very large datasets, reducing a wealth of data to its key inherent patterns may rely increasingly also on visual approaches. Data visualisation is itself a flourishing field of interdisciplinary research, and social network visualisation has established itself as an important tool for social science and media and communication studies researchers, with important early work done especially in mapping hyperlink networks between Websites in general and blogs in particular, before attention turned also to the study of network structures in more recent social networking sites. Much of this early work focussed on networks of political communication, in an effort to study the connections between contemporary political events and debates and their reflections in online environments: so, for example, Adamic & Glance (2005) and Ackland (2005) were amongst the researchers to examine the polarisation of the American political blogosphere during the 2004 U.S. presidential election season, while Kelly & Etling (2008) developed a comprehensive snapshot of politically and culturally focussed clusters of bloggers in the Iranian blogosphere ahead of the disputed 2009 election (an effort which has been repeated

for a number of other national blogospheres as well; see e.g. Bruns *et al*., 2011). Park & Thelwall (e.g. 2008) have assembled a substantial body of work on patterns of interlinkage not just between bloggers, but also between the Websites of politicians and political parties in South Korea.

More recent work has broadened the focus from studies of links between Websites (and especially blogs), and now also incorporates analyses of user interactions on and across a variety of social media Websites, at various levels of scale and with attention to a range of themes from political to phatic communication. Java *et al*. (2007) and Huberman *et al*. (2008) published some of the earliest studies which examined social network structures on *Twitter* (as well as developing a number of experimental metrics to describe the structure of the network and the level of interaction between individual users), while the most recent and most comprehensive visualisations of *Twitter* and other social network interactions at present mainly appear to be available online, but not yet published as fully refereed scholarly publications (see e.g. *Twitter*, 2011, for visualisations of the traffic flows following the earthquake and tsunami in Japan).

Social network visualisation reduces complexity and enables researchers to more easily pinpoint key participants and clusters within the network, by using a variety of metrics from social network analysis (from simple measures such as in- and outdegree to more complex metrics like centrality or eccentricity) as the foundations for their visualisations. Inherent in such approaches, however, is also a danger that apparently insightful images are allowed to overshadow close quantitative as well as qualitative analysis, and that the tools for visualising networks are treated unreflexively as mere 'black boxes' converting data into graphs; instead, it is important to develop a clear understanding of the implications resulting from using one or another of the various software tools available for network visualisation, or indeed one or another of the different visualisation algorithms which such software packages may offer (to say nothing of the settings parameters for such algorithms, which will further change the shape of the resulting visuals).

There is, of course, no one universal solution to these challenges; rather, what researchers must do is develop and document approaches to social networking data analysis and visualisation which are appropriate to their specific research projects. The present paper documents one such approach, which addresses the following question: how may we visualise the emergence, interaction, and subsequent dissolution of a discursive community of *Twitter* users forming *ad hoc* around a shared #hashtag? In keeping with this question, its approach is largely focussed on providing a detailed documentation and discussion of methodological questions; further implications of this approach to researching communicative practices in online social media are discussed in the conclusion.

Tools for Mapping *Twitter* Networks

The first and most obvious challenge for *Twitter* researchers using quantitative methodologies is to gain access to their intended datasets, of course; as noted, recent access policy changes by *Twitter* and the attendant shift towards Gnip as the preferred commercial provider for *Twitter* data have further complicated matters. However, due to the structure of the *Twitter* API, this shift affects projects which aim to capture the activities of a (potentially very large) number of known *Twitter* users far more thoroughly than those which focus on one or more chosen *Twitter* #hashtags, and it remains possible to generate some potentially very large *Twitter* datasets of public tweets that contain specific #hashtags – or indeed, as noted above, keywords event without the '#' symbol (our tracking of the keyword 'tsunami' – without the '#' – since the Japanese earthquake of 11 March 2011 captured nearly 4 million public tweets in the three weeks following the disaster, for example).

The preferred tool for capturing #hashtag or keyword tweets in recent times has been the Web service *Twapperkeeper* (*TK*), which enabled its users to lodge requests for specific terms to be tracked, and then generated comprehensive archives of captured tweets which could be downloaded from the *Twapperkeeper* site in a variety of formats (both by the original archive initiator as well as by other users). However, *Twitter*'s recent policy about-face has very directly affected the site's utility for researchers: the site is no longer permitted to make tweet datasets available for download, and now only provides functionality for viewing archived tweets on screen and generating overall statistics describing the archive contents.

At the same time, an open source version of the *Twapperkeeper* platform, *yourTwapperkeeper* (*yTK*), has also been made available by the *Twapperkeeper* developers, providing very similar functionality to the original Website but not suffering from the same export restrictions (*yourTwapperkeeper* is intended for internal use by researchers rather than for the provision of public tweet archiving services as they are available on the original Website). For researchers operating in this area, then, currently the preferred solution for capturing and archiving public tweets on specific topics is to deploy a *yTK* installation for purely internal use. (*yourTwapperkeeper* is likely to require some minor modifications to operate in an institutional network environment and to generate data in a format that is immediately compatible with that of the original *Twapperkeeper* Website, but it is outside the scope of this article to discuss these changes; readers are encouraged to engage in the *yTK* open source community or to contact the authors for advice.)

Data generated by *TK*/*yTK* is available in a variety of formats, of which the comma-separated value (CSV) or tab-separated value (TSV) exports are the most useful for the purposes of our intended network analysis and visualisation. Contained in these datasets are the following fields:

- **text:** contents of the tweet itself, in 140 characters or less
- **to_user_id:** numerical ID of the tweet recipient (for @replies)
  (not always set even for tweets containing @replies)
- **from_user:** screen name of the tweet sender
- **id:** numerical ID of the tweet itself
- **from_user_id:** numerical ID of the tweet sender
- **iso_language_code:** language code (e.g. en, de, fr, ...) of the tweet sender's default language
  (not necessarily matching the language of the tweet itself)
- **source:** name or URL of the tool used for tweeting (e.g. Web, Tweetdeck, ...)
- **profile_image_url:** URL of the tweet sender's profile picture
- **geo_type:** form in which the tweet sender's geographical coordinates are provided
  (but only a very small percentage of tweets actually provide coordinates)
- **geo_coordinates_0:** first element of the geographical coordinates
- **geo_coordinates_1:** second element of the geographical coordinates
- **created_at:** tweet timestamp in human-readable format
  (set by the tweeting client – inconsistent formatting)
- **time:** tweet timestamp as a numerical Unix timestamp

(By default, *yourTwapperkeeper* data sets also include a further 'archivesource' column to indicate how tweets were retrieved; since *yTK* retrieves all of its data through *Twitter*'s streaming API, however, this information is of very limited interest. The additional column can easily be removed from the data exports in order to make the *yTK* format entirely compatible to that of *Twapperkeeper*, and the following discussion assumes that this has been done for all datasets.)

TK/*yTK* datasets of tweeting activity around specific #hashtags and keywords, then, contain all the information required for engaging in a network analysis of @reply user conversations taking place within the #hashtag community (as well as providing the basis for a large number of other possible approaches to examining these data, which cannot be addressed in the present article; see Bruns & Burgess (2011a) for a discussion of further research methods building on such datasets). The next step after capturing the #hashtag data, therefore, is processing them. Our preferred tool for processing *TK*/*yTK* datasets – once they have been exported as CSV/TSV files – is Gawk, a GNU command-line tool (available in ported versions for Windows and Mac) that implements a simple scripting language for processing CSV/TSV files (Gawk, 2011).

Gawk operates mainly through the use of regular expressions – a standard syntax for expressing character matching and filtering conditions (see e.g. Borsodi, 2000, for an introduction). Our immediate challenge in analysing *Twitter* @reply networks, for example, is to identify where the tweets we have captured contain @replies, and to whom they are directed; since the *to_user_id* field in the archives generated by our tools is

generally unreliable, we must parse each tweet itself to identify any content that matches the '*@[username]*' format for acceptable values of *[username]*. A regular expression which performs this matching can be expressed as

```
/@([A-Za-z0-9_]+)/
```

since valid *Twitter* usernames can only contain the letters A-Z (in both upper and lower case), the digits 0-9, and the underscore character '_'.

A simple Gawk script for extracting all @replies from our overall *Twitter* data would therefore use this regular expression to identify only those lines of the CSV/TSV datafiles which match this condition. However, this would not yet go far enough to generate data which is suitable for network analysis. Instead, what is necessary is that we identify all network edges in our data: that is, all available connections from an @reply sender to an @reply receiver. A single tweet can contain @replies to two or even more *Twitter* users – as exemplified in the extreme by the popular #FF (Follow Friday) meme of tipping one's hat to a collection of interesting friends:

```
#FF Great #SNA people: @jeanburgess @coffee001 @timhighfield @halmoe @snurb_dot_info
```

A single tweet of this form would result in separate network edges from the sender to each of the five @reply recipients mentioned here, for example.

To use Gawk to generate some simple @reply network data (in the form "from user > to user") for the entire dataset, a script such as `atextractfromtoonly.awk` (Bruns & Burgess, 2011b) is sufficient. The following Gawk command line is used to execute it:

```
gawk –F , -f atextractfromtoonly.awk input.csv >output.csv
```

(or replacing the `-F ,` argument with `-F \t` if the input file exists in TSV rather than CSV format); faced with the tweet above, from user *twitteruser*, it would generate the following output (note that the script also converts all usernames to lower case, in order to avoid potential problems which could occur at a later stage if data processing tools misunderstand different capitalisations of the same username to be separate entities:

```
from,to
twitteruser,jeanburgess
twitteruser,coffee001
twitteruser,timhighfield
twitteruser,halmoe
twitteruser,snurb_dot_info
```

In spite of the very simplistic network edges data structure that this approach produces, it can already be used to generate some significant insights into the overall structure of the @reply networks under observation. In order to do so, we introduce our third key tool for *Twitter* @reply network mapping: the open source network visualisation software Gephi (Gephi, 2011). While a number of other network visualisation software packages – both open source and commercial – are also available at present, Gephi's active and highly responsive open source development community, and its focus especially also on dynamic network visualisation (which we will discuss later), positions it as the most appropriate tool for our present purposes.

Gephi is able to import a simple CSV edge list as produced by the script we have introduced above, and to visualise the network described by these data in a variety of customisable formats. (It should be noted in this context that it is important to declare the network data to describe a *directed* – rather than undirected – network during the import process, since @replies made by one user to another do not mean that the recipient will necessarily @reply back; an @reply constitutes a directed, uni-directional edge in the network,

therefore.) Fig. 1 shows the results of one such network visualisation, for example, using as its source dataset the @reply data from discussion of a purported leadership challenge against then-Australian Prime Minister Kevin Rudd by his Deputy Julia Gillard, under the *Twitter* hashtag #spill (Australian politicial slang for such a challenge) in the evening of 23 June 2010:

[insert fig. 1 here]

Fig. 1: *#spill @reply network, 23 June 2010. Node size = indegree; node colour = outdegree.*

While it is not the purpose of this article to discuss the findings of our research into patterns of participation in the #spill discussion, some brief notes on Fig. 1 will help readers understand the potential utility of such network analysis and visualisation approaches (see Bruns & Burgess, 2010, for a more detailed discussion of *Twitter* use during the leadership spill and the subsequent federal election). First (with node size set to indicate the amount of @replies received), it is obvious that a large number of participants in the #spill conversation not only discuss the potential fate of the Prime Minister, but do so while referring to him using his *Twitter* username @KevinRuddPM rather than merely his name. At the same time (and unsurprisingly), this significant level of incoming @replies does not result in any responses *from* the Prime Ministerial account; with node colour indicating the amount of @replies sent, @KevinRuddPM remains a pale yellow, indicating no activity. Conversely, there are also a number of very highly active senders of @replies (shown in red), who do not necessarily also receive a significant number of answers to their messages – and most interestingly, perhaps, it is possible for us to identify a handful of participants who are notable *both* as senders *and* as recipients of @replies (acting as central hubs in the network); most notably, these include journalists @latikambourke and @renailemay (indicating, incidentally, the continued importance of mainstream media sources even in social media environments).

It should be noted in this context that our approach so far makes no difference between simple @replies as they are used to conduct a conversation between two or more *Twitter* users, and a specific form of @reply known as a (manual) retweet:

```
RT @GreenJ: Newspaper correction of the year. The Sun. Winning. http://bit.ly/SQ7Ms
```

In this manual form (`RT @[username] [original tweet]`), retweets contain an @reply preceded by the abbreviation 'RT'; while it would be easily possible to distinguish them from other @replies which do not contain the 'RT', therefore, we choose not to do so at this point because manual retweets often also serve a conservational purpose, as retweeters frequently use their retweet to comment on what they are sharing:

```
OK. This is getting silly. RT @Telegraph: Welsh harpist ready for Royal Wedding
http://tgr.ph/fXw62f
```

By contrast, *Twitter*'s more recently introduced 'retweet button' functionality can only be used to share other users' tweets verbatim; it simply inserts the original tweet in the retweeting user's tweet timeline, and does not add a '`RT @[username]`' to the retweeted content. Such new-style 'button' retweets are not captured by *Twapperkeeper* or *yourTwapperkeeper*, however, and are therefore outside the scope of what our methods can address.

In our analysis, then, which is therefore concerned only with standard @replies and old-style '`RT @[username]`' retweets as a specific form of @reply, it will at times be important to distinguish between these two types; a high number of (manual, old-style) retweets of a salient tweet can substantially boost the indegree (@replies received) count of the originating *Twitter* user, for example, and the inclusion of such outliers may or may not be desirable in specific research contexts. This is a case-by-case choice which cannot be resolved in this more general outline of our methodology, however; for our present purposes, suffice to say

that it is easily possible to distinguish @replies from manual retweets in the data, and to eliminate one or the other category from the dataset if necessary.

Towards Dynamic @Reply Visualisations

Static network analyses and visualisations as we have been able to create them with the methodological approaches introduced so far are rarely an end in themselves, it should be noted; rather, they can be useful complemented with further qualitative and quantitative analysis of tweet content, investigations of the actions of individual users (for example those which represent notable authorities and institutions), and other research. Indeed, as we have seen even from our brief discussion of the #spill example, the network analysis itself can be used to pinpoint those users whose activities it may be most interesting to study in further detail.

In addition to these rich research opportunities, however, it is also possible to do yet more advanced research work with the original datasets themselves, especially once we begin to utilise the timeline data which is also available to us. As we have seen above, each tweet in our dataset is associated with a specific timestamp in both human-readable and numeric formats, and these timestamps hold the key to a large number of further opportunities for processing the data. Our eventual goal in this effort is to develop what we can describe as dynamic network visualisations: network maps which are time-dependent and change as the specific timeframe under observation changes; to arrive at this point, however, it is first useful to examine the overall temporal patterns which exist in the data.

A first step in this process is to correctly identify the exact time at which specific tweets were sent. While the data captured by *Twapperkeeper* or *yourTwapperkeeper* already contains a human-readable timestamp, it is immediately obvious from a brief glance at the '`created_at`' field that the format in which the tweet time is provided here is inconsistent: although the majority of times are given in the format 'Sat Jun 19 02:15:03 +0000 2010', a small percentage is stated as 'Sat 19 Jun 2010 02:17:07 +0000' or in other variations on the theme. However, the *TK*/*yTK* data structure also provides a second timestamp field, '`time`', which contains the unique Unix timestamp of the tweet – a number indicating the seconds which have elapsed since 00:00 on 1 January 1970, UTC. As this figure is unique, and unaffected by timezones or time shifts due to daylight savings or similar schemes, it provides an ideal basis for a Gawk script which converts the timestamp into a variety of formats of interest: `explodetime.awk` (Bruns & Burgess, 2011b).

This script removes the original two time-related fields of the archive (the human-readable '`created_at`' and the Unix timestamp '`time`'), and adds the following eight fields in their stead:

- **time:** full human-readable timestamp
- **timestamp:** full Unix timestamp
- **condensed:** Unix timestamp, condensed to 100-second intervals (i.e. last two digits changed to '00')
- **year:** YYYY only
- **month:** YYYY-Mon
- **day:** YYYY-Mon-DD
- **hour:** YYYY-Mon-DD HH
- **minute:** YYYY-Mon-DD HH:MM

The script would be run using Gawk as follows:

```
gawk -F , -f explodetime.awk input.csv >output.csv
```

(or again with the argument `-F \t` rather than `-f ,` if the input file is in TSV rather than CSV format). It should be noted that the resulting human-readable timestamps of various formats which are provided in the added fields are always given by default in the researcher's own timezone (that is, in the timezone of the machine on which Gawk is run); this behaviour can be changed by modifying the `explodetime.awk` script to add or

subtract a set number of seconds from the original timestamp (e.g. adding 5 x 60 x 60 = 18,000 seconds to shift all times five hours into the future), or more simply by changing the local machine's timezone before the Gawk script is executed.

The added fields can then be used to graph the total volume of tweets against a timeline, at various levels of resolution: per year (for very long-term datasets), month, day, hour, or minute, or per 100-second interval by using the 'condensed' field. How this is done in practice depends on the specific software used, and need not concern us here; in Excel, for example, it would simply be a matter of inserting a Pivot Table and graphing a time interval (e.g. 'hour') against the number of tweets which share the same hour-level timestamp ('count of text'). The hour-by-hour activity in the #spill hashtag during 23 June 2010 (AEST), for example, indicates that first rumours of a potential leadership challenge for the Prime Ministership only emerged at some point after 19:00 that night, peaking between 22:00 and 23:00 (at over 13,400 tweets per hour, or over 220 tweets per minute) when it was confirmed that a parliamentary Labor Party caucus vote would indeed be held the following day.

Such indications of activity patterns over time are valuable in their own right, of course – and they can be further extended by filtering the overall tweet data for specific keywords before graphing the specific rise and fall of such keywords against one another over time, for example (see Bruns, 2010a/b/c/d and 2011a/b/c for analyses of *Twitter* use in the 2010 Australian election and the 2011 Christchurch earthquake which take this approach, for example). In the context of our present discussion, however, they are also simply useful to pinpoint the exact timeframe within the full dataset for which we will want to develop a dynamic @reply visualisation; in the case of the #spill data, for example, this would clearly be the five-hour period between 19:00 and midnight on 23 June 2010 (AEST).

First, then, we must filter our dataset using the `timeframe.awk` helper script (Bruns & Burgess, 2011b), which takes a 'start' and 'end' argument, each in the format 'YYYY MM DD HH MM SS' (again assumed to be in the local user's timezone), to define the timeframe to select from the overall data, and is executed as follows:

```
gawk -F , -f timeframe.awk start="2010 06 23 19 00 00" end="2010 06 24 00 00 00"
input.csv >output.csv
```

(using the 19:00 to 00:00 #spill timeframe highlighted above for this example). The output of this script is a subset of the original dataset which includes only those tweets that fall into the selected timeframe.

The next steps begin the process of preparing our dataset in preparation for dynamic network visualisation. Instead of the simple `atextract.awk` script which we used to visualise the overall network of @replies regardless of when they were made, we must now find a way to preserve not just information about which user sent an @reply to whom, but also the point in time when each @reply was made; ultimately, these data need to be presented in a format which is intelligible to Gephi. That format is the GEXF standard (GEXF, 2011), an XML-based format introduced by Gephi.

To generate time-based network structure data from *Twapperkeeper*/*yourTwapperkeeper* datasets requires two steps, and two separate Gawk scripts. First, we process the *TK*/*yTK* data using the `preparegexfattimeintervals.awk` script (Bruns & Burgess, 2011b), which – similar to `atextract.awk` – identifies all @replies in our data, but collates them for each 'from > to' pair and also records the specific timestamps at which the @replies from one user to another were made. A sample line of output from this script may read:

```
from,to,timestamps
snurb_dot_info,jeanburgess,10;245;3452
```

which would indicate that @snurb_dot_info sent @replies to @jeanburgess at three points: 10 seconds, 245 seconds, and 3452 seconds after the start of the timeframe covered in the dataset. (Note that, to aid usability, the timestamps generated by this script no longer indicate Unix time, therefore, but simply count the seconds since the first tweet that is included in the dataset; alternatively, the start time which is to be taken as $t = 0$ can

be set manually to a specific Unix timestamp using the 'offset' argument on the command line.) The script is invoked as follows, then:

```
gawk -F , -f preparegexfattimeintervals.awk input.csv >output.csv
```

or, if a specific start time is to be used,

```
gawk -F , -f preparegexfattimeintervals.awk offset="1277283600" input.csv
>output.csv
```

(again using the example of 19:00 AEST on 23 June 2010, which translates to a Unix timecode of 1277283600).

How Long Is a Tweet?

A second script, `gexfattimeintervals.awk` (Bruns & Burgess, 2011b), then converts the output of `preparegexfattimeintervals.awk` (which still takes the form of a CSV or TSV file) into the GEXF format. In doing so, however, we begin to encounter the fundamental question which lends this article its title: how long is a tweet? From one perspective, the spatial, this question has an obvious answer: a tweet is up to 140 characters long. In approaching *Twitter* and its conversational processes as a dynamic space, however, we encounter tweets in a second, and more interesting, dimension: the temporal. From this perspective, the question 'how long is a tweet' comes to mean 'how long does a tweet – or in the present case, more precisely, an @reply – last?'

In visualising the dynamic network created through the practice of @replying, the different answers to this question result in vastly different perspectives on the network. In the first place, we might suggest that an @reply, once made, simply *persists*: with each @reply (including manual retweets, for the purposes of our present discussion), a new connection between two users is made, or an existing one strengthened further; starting from a blank space at the time zero which we have chosen for our dataset, a network emerges which grows progressively denser, and whose individual edges grow gradually thicker as individual users repeatedly @reply to each other. This maximal perspective, from which an @reply connection from one user to another, once made, never disappears again, lends itself well to the study of #hashtags as they emerge, in fact: it provides a way to show how more and more users join the #hashtag conversation and how the network thickens as a process.

Conversely, a minimal perspective, which focusses more strongly on the dynamics of *Twitter* interaction, would treat individual @replies as highly ephemeral, and meaningful only in cumulative form: from this perspective, we could argue – in line with long-established observations about the ephemerality of online spaces (see e.g. MacKinnon, 1995) – that individual @replies are visible to sender and recipient only very briefly, and meaningful only if the connection between the two is upheld through frequent ongoing exchanges; a single @reply constitutes only the most fleeting of connections, therefore, and can be said to have disappeared within seconds. A visualisation which takes this approach would not generate a strongly and increasingly interconnected cumulative network map, but rather a rapidly changing map where the centre of activities shifts quickly across the network, depending on which participants are tweeting at one another at any one point. This, too, is a valuable perspective, enabling researchers to examine how sections of the network react and respond to specific events (for example to new information coming to hand and being shared by *Twitter* users).

In between the maximalist and minimalist extremes there exists a sliding scale of other possibilities, according @replies a greater or smaller lifetime. The obvious analogy here is to the half-time of radioactive elements, which measures how long it takes an element to decay by emitting its constituent subatomic particles. Here, we similarly measure the decay time of a tweet – that is, the time it takes until a previous @reply should no longer be considered part of the current #hashtag conversation network. Contrary to particle physics, there is no specific metric that would provide an exact value for this decay time; the choice of

value is simply a matter for the researcher themselves. However, a number of contextual factors may need to be considered in each specific case:

- First, the overall volume of tweets per minute which presently occur within the #hashtag conversation should be considered: the more active a #hashtag is overall, the more quickly will previous tweets (including @replies) made within the #hashtag conversation scroll out of sight for an individual *Twitter* user following the #hashtag. @replies in a slow-moving #hashtag conversation may be seen to have a longer decay time than those in a very active #hashtag group.
- Second, the number of @replies received by the same user during any one timeframe similarly affects how quickly individual @replies are buried under subsequent tweets. @replies made to a very 'popular' user may be seen to decay more quickly than those to a user with fewer correspondents.
- Third, beyond the specific #hashtag itself, @replies to *Twitter* members who follow a larger number of other users may not be visible to them for as long as @replies to relatively poorly-connected users; @replies to more omnivorous *Twitter* members may be seen to decay more quickly than those to more selective participants, therefore.
- Fourth, we might also want to take into account whether the two users connected by an @reply are already known to one another (and perhaps follow one another), or whether they have encountered one another for the first time – which is a significant possibility especially when examining *Twitter* activity around a #hashtag community, brought together as it is by a common topic rather than existing follower/followee structures. An @reply between two users who already follow one another could very sensibly be seen to be more meaningful than one between two users who have never heard of one another before.

At the same time, unless we have access to data well beyond what is available from standard *Twapperkeeper/yourTwapperkeeper* archives, we will usually be able to make an informed judgment only on the first (and to a limited extent perhaps also the second) of these factors; further, to set an @reply decay time which takes into account the personal circumstances of each individual participant (which is what a full consideration of points two to four would require) would introduce considerable – possibly unmanageable – complexity.

What we *are* able to do relatively straightforwardly, then, is simply to select a global decay time for @replies, with that decay time representing one of three possible perspectives: the maximalist (decay time approaching infinity – @replies persist indefinitely), the minimalist (decay time close to zero – @replies decay immediately), and the intermediate (decay time based on contextual factors – @replies decay after a reasonable time). For the context of the #spill example which we have discussed above, where we are examining a five-hour period between 19:00 and midnight, a sensible intermediate decay time may range between 15 and 30 minutes, for example.

Armed with these possible answers to the question 'how long is a tweet', we are now able to use `gexfattimeintervals.awk` to generate a GEXF-format representation of a dynamic *Twitter* #hashtag @reply network. The script takes an optional `decaytime` argument, which specifies the global @reply decay time in seconds (the decay time is set to 100 seconds by default if no such argument is provided). It takes as its input the intermediary output generated by `preparegexfatintervals.awk`, and is called as follows:

```
gawk -F , -f gexfattimeintervals.awk decaytime=1 input.csv >output.gexf
        (minimalist option – @replies decay after 1s)
gawk -F , -f gexfattimeintervals.awk decaytime=1800 input.csv >output.gexf
        (intermediate option – @replies decay after 1800s = 30mins)
gawk -F , -f gexfattimeintervals.awk decaytime=31536000 input.csv >output.gexf
        (maximalist option – @replies decay after one year)
```

(note that the output file now takes the .gexf extension, of course; it is no longer a CSV file).

This GEXF file can now be loaded in to Gephi for network visualisation; while the GEXF file itself contains only data on the edges in the network (including the starting and ending nodes of the connection, and their various start and end timecodes), and not on the nodes themselves, on loading the file Gephi will immediately generate the node information itself. This also means that the nodes themselves will be visible throughout the entire timeframe, disconnected from the existing network during those times when no edges are active; in future revisions of the scripts presented here, we expect to add functionality that will enable the nodes themselves to disappear during the times that they are inactive.

Using the timeframe selector slider in the Gephi window, it is now possible to choose specific timeframes for visualisation. Depending on the total size of the network, it may be advisable first to filter the overall network to include only those nodes which are the most connected, as standard desktop machines will struggle to cope with the visualisation of networks that contain well over 1000 nodes. To do so, the following steps need to be followed:

1. Load the full GEXF file into Gephi (File menu > Open).
2. Run the Average Degree statistics measure (Statistics tab > Network Overview).
3. Filter for nodes with indegree > 10 (Filters tab > Attributes > Range > Degree; set range minimum to 10; click Filter).
4. Select and copy entire visible graph (Graph tab > Rectangle selection tool > select graph; Right-click on graph and Copy to > New workspace).
5. Switch to new workspace (click Workspace 0 in the bottom right-hand corner of the Gephi window, select Workspace 1).
6. Export filtered data to new GEXF file (File menu > Export > Graph file; select GEXF format and save).
7. Close and reopen Gephi, and load exported network data.

(The exact indegree cutoff value in step 3. may be varied depending on the desired total size of the network, of course.)

The next choice we must confront in visualising the dynamic @reply network within a #hashtag community concerns form of network dynamics which we wish to see. One option is to begin by using Gephi to visualise the network for the entire timeframe covered in the GEXF file; this generates a cumulative network map that takes into account all @replies, regardless of the point in time at which they were made, and – depending on the specific algorithms and settings chosen for the visualisation itself – positioning the network nodes (i.e. the participating *Twitter* members) according to their overall participation in the @reply network across the entire time period. Once the visualisation process concludes, and the visualisation algorithm has terminated, we may now use the timeframe selector to display only those connections between nodes on the network map which were active within the selected timeframe, without changing the positioning of individual nodes at all. In other words, selecting a timeframe in this way only shows or hides a subset of the connections between individual nodes, while the nodes themselves remain stationary. An anatomic analogy for this form of visualisation would be the brain scan, which superimposes on the established and largely immobile structure of the brain itself a visualisation of where in this structure activity is currently taking place.

Alternatively, we may begin by first selecting a specific timeframe (for example, the first 15 minutes) of the overall time period represented in the data, and then starting the network visualisation algorithm. Now, the network structure which emerges as the algorithm runs reflects only those connections which are active during the selected timeframe, with other nodes in the network floating as disconnected entities on the periphery of the network graph. As the timeframe selection changes (by moving the slider forwards or backwards), new nodes may join the network as @replies made during the new timeframe become active, while others leave the network if the @replies which connected them have passed their point of decay. (Note in this context that a number of the available visualisation algorithms in Gephi take some time to update the

graph; a rapid movement of the timeframe slider will not produce the intended results, therefore.) This form of visualisation highlights the dynamic nature of the conversation more strongly, and shows more clearly the potential shifts both in the overall level of participation and in who the most central participants may be at any one time (for that reason, it *may* therefore also be more appropriate in visualising a #hashtag network).

[insert fig. 2a/b here]

Fig. 2a: *#spill network – node positions fixed* Fig. 2b: *#spill network – node positions variable*
(algorithm: Force Atlas; node size: indegree; node colour: degree; timeframe: 19:30-20:00, 23 June 2010)

Fig. 2 provides two snapshots of the dynamic network visualisations which result from this process, demonstrating the differences between these two approaches to visualising the network: while both graphs depict the same timeframe within the overall #spill dataset (the period between 19:30 and 20:00 on 23 June 2010, when rumours of a leadership challenge first surfaced in earnest), in Fig. 2a all nodes in the network are placed in fixed positions according to their overall @reply patterns during the entire period from 19:00 to midnight, but only the @replies active during the currently chosen timeframe are shown as connections between them; in Fig. 2b the positioning of nodes is itself subject to which @replies are currently active, and can therefore change over time (for the 19:30 to 20:00 period, this results in an interconnected network core in the top left quadrant, and a pool of disconnected nodes which are not currently participating in any @reply conversations). In both versions, node size indicates indegree (the number of @replies received, over the entire five-hour period), and node colour indicates degree (the number of @replies sent or received over that period). For animated visualisations of the #spill network data that further demonstrate the differences between these two modes of dynamic network visualisation, see Bruns (2010e/f).

Neither of these visualisation options is inherently more or less 'correct', of course; the choice between the two depends largely on what aspects of the network dynamics researchers intend to highlight. A further choice also exists in the exact size of the timeframe (as a subset of the total time period covered by the dataset) which is selected for visualisation: in creating a dynamic visualisation using Gephi, the timeline slider simply allows users to set specific start and end points (between the total start and end times of the overall dataset). The exact size of the chosen subset of time – a period which we might call the 'time aperture', as it defines the subset of the total network data which is visualised at any given point – interacts in significant ways with the @reply decay time which we have discussed above.

Gephi includes in its network visualisation any nodes or edges whose period of existence overlaps with the chosen time aperture. Assuming that we have set an @reply decay time of 30 minutes (1800s) in generating the GEXF data, and that we have also chosen a time aperture of 30 minutes, then, the total network included in Gephi's graph will stretch over a timeframe of up to 90 minutes, therefore: an @reply made just under 30 minutes before the start of the chose time aperture would still be included since it would not have fully decayed by the time the aperture opens, while a tweet made in the dying seconds of the time aperture would not decay for another 30 minutes beyond its closing. Again with an @reply decay time of 30 minutes, even a theoretically possible time aperture of no more than a few seconds could still result in a network graph that covers a total period of nearly one hour, as it would include any @replies that expire during the aperture period (and were therefore active during the preceding 30 minutes), any @replies which started before and finish after the aperture period, and any @replies which are made during the aperture period (and expire 30 minutes later). This, then, needs to be remembered in choosing the time aperture to be visualised in Gephi – and unless the decay time itself is already relatively minimal, it is generally advisable to keep the time aperture itself short if the aim is to visualise relatively brief periods in the data.

Conclusion: Uses for Dynamic *Twitter* Network Visualisation

The dynamic network visualisations which our approach enables us to generate are significant in their own right; depending on the nature of the #hashtag data to be visualised, and on the period for which data is

available, they enable us to highlight the shifting roles played by individual participants over time, as well as the response of the overall #hashtag community to new stimuli – such as the entry of new participants or the availability of new information. Over longer timeframes, it may be possible to identify different phases in the overall discussion, or the formation of distinct clusters of preferentially interacting participants. Such observations may also be combined with other approaches that study the prevalence of specific themes and topics within the #hashtag community, track the parallel development of mainstream media coverage of the event being studied, or correlate interactions within the #hashtag community with underlying follower/followee structures (to name just a few possibilities).

Additionally, of course, there are many more possibilities for visualising these dynamic data. Gephi itself already provides a substantial number of alternative network visualisation algorithms, whose outputs can be further modified using a range of specific settings; which algorithms and what settings lend themselves best to the visualisation of #hashtag @reply data, in pursuit of specific research questions, is a significant area for further enquiry that is, however, well beyond the scope of the present paper. Similarly, beyond mere measures of indegree and outdegree (received and sent @replies, in our case), as we have used them in the network map presented in Fig. 1, a variety of other metrics are also available to pinpoint and highlight nodes of interest within the network; which of these are relevant to the study of #hashtag communities is, again, a matter for further exploration beyond our present discussion.

What a solely #hashtag-based approach to the study of *Twitter* interactions does not enable us to examine, by contrast, is the level of relevant interaction that may take place outside the #hashtag proper, or under other, alternative #hashtags. This is especially relevant for the study of @replies, as we have presented it here: not all @replies which result from interactions in the #hashtag community will themselves also again include the #hashtag; indeed, those @replies which deliberately do include the #hashtag (in order to make themselves visible to the wider community) may be said to be engaged in a public *performance* of conversation as much as they are engaged in the conversation itself. A further extension of our approach – which especially under *Twitter*'s current, more restrictive interpretation of its rules for providing data access to researchers will be substantially more difficult, however – would therefore not only take into account all the tweets using the given #hashtag itself, but would also seek to capture all follow-on tweets by #hashtag participants, even if those tweets themselves no longer contain the #hashtag.

Significant additional opportunities for further research also exist in the direction of analyses which combine the data on conversational exchanges with available indicators of thematic content, or other markers. For example, when do new themes emerge in the #hashtag community; how and to what extent do they flow across the network? In particular, how are URLs and other references to external media shared, and what impact do they have on conversational exchanges? In larger #hashtag communities, what role do existing user attributes play: do exchanges about international themes separate into various communicative groups defined by shared language or geography? Does it matter what *Twitter* clients are used (desktop or mobile, Website or stand-alone application)? To what extent do pre-existing follower/followee networks predetermine who responds to whom? The diachronic, dynamic dimension of mapping which we have explored here is able to provide valuable additional perspectives on each of these questions: so, for example, we might test the assumption that conversations take place along established follower/followee connections more strongly during earlier than during later stages in the #hashtag's lifetime, or test whether geographically or thematically 'local' users are both the earliest and the most persistent participants in #hashtag conversations around a major breaking story.

Similarly, how #hashtags emerge as coordinating mechanisms for conversations between groups of *Twitter* users with related interests also remains to be explored in more detail. In some cases, specific #hashtags are announced or suggested through other means (for example by conference organisers promoting *Twitter* use by delegates, or by television channels hoping to create additional social media buzz around their shows); in many other cases, however, they emerge more gradually as users tweeting about similar topics find one another and settle on a common way of tagging their tweets. *Mutatis mutandis*, the dynamics of arriving at a shared approach to #hashtagging specific conversations may well be comparable to the processes by which

other open, crowd-driven tagging systems settle into stable patterns (see e.g. Golder & Huberman, 2005, for an investigation into the converging dynamics of thematic tags assigned to URLs shared through the social bookmarking site *del.icio.us*).

Finally, it must also be noted that not all #hashtags are alike. While we have focussed here on hashtags which are thematically defined, and possibly focussed around current events or breaking news, a wide range of other #hashtag uses also exist: from marking long-established, recurring *Twitter*-based get-togethers of likeminded users (such as #phdchat, a weekly gathering of PhD students for social and scholarly support) through coordinating short-lived, often humorous *Twitter* memes (such as #tweetlikemarlonbrando, encouraging users to impersonate well-known celebrities) to serving as discursive markers emphasising the *Twitter* user's views (such as #fail, #facepalm, #headdesk, and many more). It would be an exaggeration to speak of the existence of #hashtag *communities* in each of these cases – but this does not invalidate the research methods outlined here: indeed, these methods will serve to highlight the comparative absence of conversational networks in those cases where #hashtags are used mainly as exclamation points, or to mark tweets as contributing to a specific meme.

Such further opportunities for research, then, also point again to the validity of Rogers's call to "follow the medium" (2009: 10): especially given the continuing, sometimes rapid, evolution of *Twitter* as a medium, and of its users' practices of communication, it remains necessary for researchers to allow the patterns emerging from their data to direct the focus of their further work at least to some extent. The central purpose of this paper, then, is not to present our methodology for the dynamic visualisation of @reply interactions in *Twitter* #hashtag communities as a final outcome, but instead to position it as a crucial enabler for further, more detailed work which combines the findings which this approach can generate with the results of a variety of other, similarly innovative means of processing *Twitter* datasets; we hope to be able to use this paper to ignite rather than to conclude an continuing conversation about more sophisticated ways of dealing with *Twitter* data. To this end, we also encourage readers to make contact with our research team through our project blog, at http://mappingonlinepublics.net/.

References

Ackland, Robert. (2005). "Mapping the U.S. political blogosphere: Are conservative bloggers more prominent?" *BlogTalk Downunder 2005*, Sydney, 19-22 May 2005. Available at: http://voson.anu.edu.au/papers/polblogs.pdf (accessed 26 July 2011).

Adamic, L. A., & Glance, N. (2005) "The political blogosphere and the 2004 U.S. election: Divided they blog." *2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, Chiba, Japan, 10 May 2005. Available at: http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf (accessed 26 July 2011).

Borsodi, Jan. (2000, 30 Oct.) "Regular Expressions explained." *Zez.org: About Code*. Available at: http://zez.org/article/articleview/11/ (accessed 1 Apr. 2011).

boyd, danah, Scott Golder, and Gilad Lotan. (2010) "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter." *HICSS-43*. IEEE: Kauai, HI, 6 January 2010. Available at: http://www.danah.org/papers/TweetTweetRetweet.pdf (accessed 26 July 2011).

Bruns, Axel. (2010a, 1 Sep.) "Trends on #ausvotes during the Australian election, pt. 1." *Mapping Online Publics*. Available at: http://mappingonlinepublics.net/2010/09/01/trends-on-ausvotes-during-the-australian-election-pt-1/ (accessed 1 Apr. 2011).

———. (2010b, 1 Sep.) "Trends on #ausvotes during the Australian election, pt. 2." *Mapping Online Publics*. Available at: http://mappingonlinepublics.net/2010/09/01/trends-on-ausvotes-during-the-australian-election-pt-2/ (accessed 1 Apr. 2011).

———. (2010c, 1 Sep.) "Trends on #ausvotes during the Australian election, pt. 3." *Mapping Online Publics*. Available at: http://mappingonlinepublics.net/2010/09/01/trends-on-ausvotes-during-the-australian-election-pt-3/ (accessed 1 Apr. 2011).

———. (2010d, 1 Sep.) "Trends on #ausvotes during the Australian election, pt. 4." *Mapping Online Publics*. Available at: http://mappingonlinepublics.net/2010/09/01/trends-on-ausvotes-during-the-australian-election-pt-4/ (accessed 1 Apr. 2011).

———. (2010e, 30 Dec.) "Visualising Twitter dynamics in Gephi, part 1." *Mapping Online Publics*. Available at: http://mappingonlinepublics.net/2010/12/30/visualising-twitter-dynamics-in-gephi-part-1/ (accessed 1 Apr. 2011).

———. (2010f, 30 Dec.) "Visualising Twitter dynamics in Gephi, part 2." *Mapping Online Publics*. Available at: http://mappingonlinepublics.net/2010/12/30/visualising-twitter-dynamics-in-gephi-part-2/ (accessed 1 Apr. 2011).

———. (2011a, 16 Mar.) "Twitter in the Christchurch earthquake, pt. 1." *Mapping Online Publics*. Available at: http://mappingonlinepublics.net/2011/03/16/twitter-in-the-christchurch-earthquake-pt-1/ (accessed 1 Apr. 2011).

———. (2011b, 16 Mar.) "Twitter in the Christchurch earthquake, pt. 2." *Mapping Online Publics*. Available at: http://mappingonlinepublics.net/2011/03/16/twitter-in-the-christchurch-earthquake-pt-2/ (accessed 1 Apr. 2011).

———. (2011c, 16 Mar.) "Twitter in the Christchurch earthquake, pt. 3." *Mapping Online Publics*. Available at: http://mappingonlinepublics.net/2011/03/16/twitter-in-the-christchurch-earthquake-pt-3/ (accessed 1 Apr. 2011).

———, and Jean Burgess. (2010) "Election 2010: The View from Twitter." Paper presented at the International Australian Studies Association 'Double Vision' conference, Sydney, 26 Nov. 2010. Available at: http://mappingonlinepublics.net/2010/11/24/election-2010-the-view-from-twitter/ (accessed 1 Apr. 2011).

———, and Jean Burgess. (2011a) *Mapping Online Publics*. Available at: http://mappingonlinepublics.net/ (accessed 1 Apr. 2011).

———, and Jean Burgess. (2011b, 22 June) "Gawk scripts for Twitter processing." v1.0. *Mapping Online Publics*. Available at: http://mappingonlinepublics.net/resources/ (accessed 26 July 2011).

———, Jean Burgess, Tim Highfield, Lars Kirchhoff, and Thomas Nicolai. (2011) "Mapping the Australian networked public sphere." *Social Science Computer Review* vol. 29, no. 3 (Aug. 2011): 277-287. Available at: http://ssc.sagepub.com/content/29/3/277.full.pdf+html (accessed 26 July 2011).

Gawk. (2011) Available at: http://www.gnu.org/software/gawk/ (accessed 1 Apr. 2011).

Gephi. (2011) Available at: http://gephi.org/ (accessed 1 Apr. 2011).

GEXF. (2011) "GEXF File Format." Available at: http://gexf.net/ (accessed 1 Apr. 2011).

Golder, Scott A., and Bernardo A. Huberman. (2005) "The structure of collaborative tagging systems." Information Dynamics Lab, HP Labs. http://arxiv.org/ftp/cs/papers/0508/ 0508082.pdf (accessed 21 Sep. 2007).

Halavais, Alexander, and Helen Martin-Elmer. (2009) "Back@you: Tracing the diffusion of a conversational convention." Paper presented at the Association of Internet Researchers conference, Milwaukee, 10 Oct. 2009.

Honeycutt, Courtenay, and Susan C. Herring. (2009) “Beyond microblogging: Conversation and collaboration via Twitter.” *Forty-Second Hawai'i International Conference on System Sciences*, Los Alamitos, CA, 2009. Available at: http://ella.slis.indiana.edu/~herring/honeycutt.herring.2009.pdf (accessed 26 July 2011).

Huberman, Bernardo, Daniel Romero, and Fang Wu. (2008) “Social networks that matter: Twitter under the microscope.” *First Monday* vol. 14, no. 1 (20 Dec. 2008). Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2317/2063 (accessed 1 Apr. 2011).

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. (2007) “Why we Twitter: Understanding microblogging usage and communities.” *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, San Jose, California, 12 Aug. 2007*. Available at: http://ebiquity.umbc.edu/get/a/publication/369.pdf (accessed 1 Apr. 2011).

Kelly, J., & Etling, B. (2008) “Mapping Iran’s online public: Politics and culture in the Persian blogosphere.” Boston: Berkman Center for Internet & Society. Available at: http://cyber.law.harvard.edu/publications/2008/Mapping_Irans_Online_Public (accessed 26 July 2011).

MacKinnon, Richard C. (1995) “Searching for the leviathan in Usenet.” In Steven G. Jones, ed., *CyberSociety: Computer-Mediated Communication and Community*. Thousand Oaks, Calif.: Sage. 112-37.

Melanson, Mike. (2011, 11 Feb.) “Twitter kills the API whitelist: What it means for developers & innovation.” *ReadWriteWeb*. Available at: http://www.readwriteweb.com/archives/twitter_kills_the_api_whitelist_what_it_means_for.php (accessed 1 Apr. 2011).

Mendoza, Marcelo, Barbara Poblete, and Carlos Castillo. (2010) “Twitter under crisis: Can we trust what we RT?” *Social Media Analytics, KDD '10 Workshops*, Washington, DC, 25 July 2010. Available at: http://research.yahoo.com/files/mendoza_poblete_castillo_2010_twitter_terremoto.pdf (accessed 26 July 2011).

Park, Han Woo, and Michael Thelwall. (2008) “Link analysis: Hyperlink patterns and social structure on politicians’ Web sites in South Korea.” *Quality and Quantity* vol. 42, no. 5 (2008): 687-697.

Rogers, Richard. (2009) *The End of the Virtual: Digital Methods*. Amsterdam: Vossiuspers UvA. Available at: http://www.govcom.org/publications/full_list/oratie_Rogers_2009_preprint.pdf (accessed 26 July 2011).

———. (2010) “Internet research: The question of method.” *Journal of Information Technology & Politics* vol. 7 (2010): 241-260. Available at: http://www.govcom.org/publications/full_list/rogers_internet_research_question_of_method_2010.pdf (accessed 26 July 2011).

Shiels, Maggie. (2011, 28 Mar.) “Twitter co-founder Jack Dorsey rejoins company.” *BBC News*. Available at: http://www.bbc.co.uk/news/business-12889048 (accessed 1 Apr. 2011).

Twapperkeeper. (2011) Available at: http://twapperkeeper.com/ (accessed 1 Apr. 2011).

Twitter. (2011, 29 June) “Global pulse.” Available at: http://blog.twitter.com/2011/06/global-pulse.html (accessed 26 July 2011).

yourTwapperkeeper. (2011). Available at: http://your.twapperkeeper.com/ (accessed 1 Apr. 2011).

Figure 1:

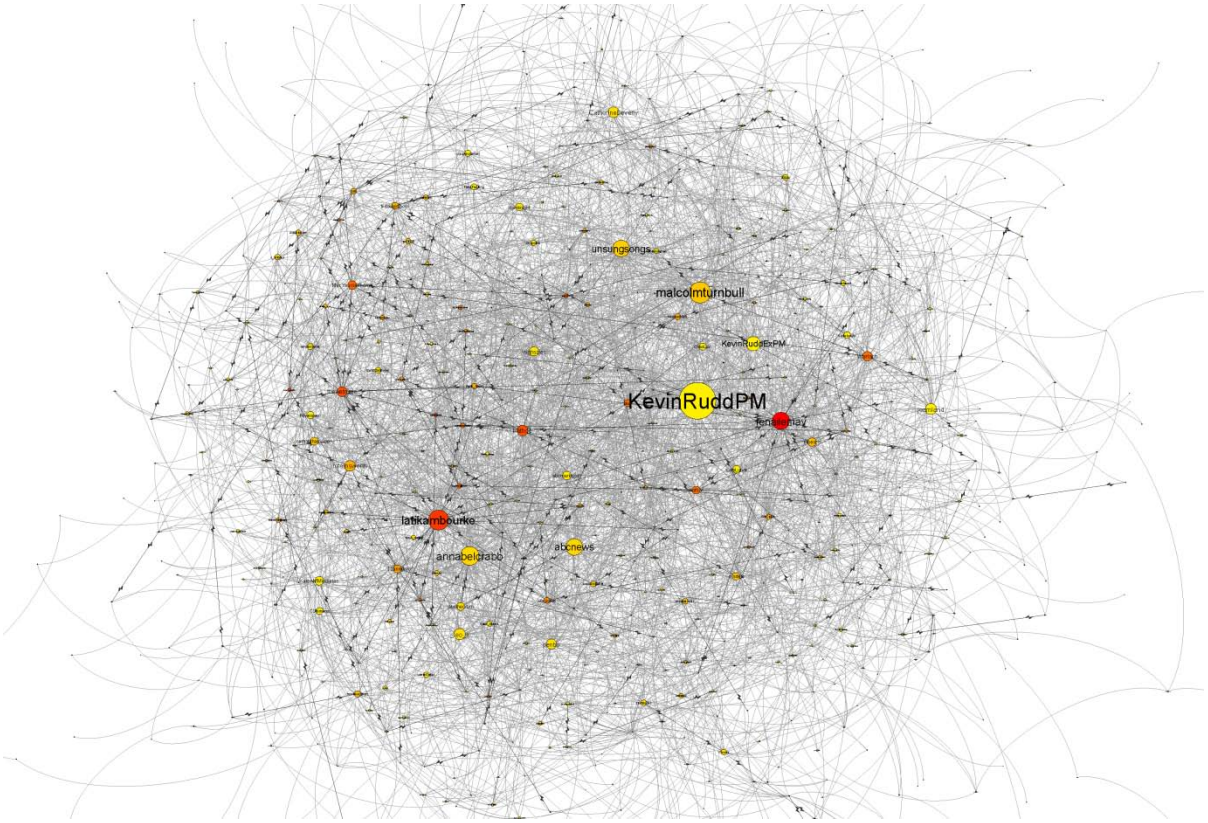

Figure 2a/b: