Analysing Twitter Data with Text Mining and Social Network Analysis

Analysing Twitter Data with Text Mining and Social Network Analysis

Yanchang Zhao

Intent Management and Analytics Section,
Risk Analysis and Monitoring Branch,
Department of Immigration and Citizenship, Australia
Email: yanchang.zhao@immi.gov.au

Abstract

Twitter, as one of major social media platforms, provides huge volume of information. A research project was performed recently in the Analytics Research Weeks in the Australian Department of Immigration and Citizenship (DIAC) to analyse Twitter data and study the feasibility of using them to better understand and improve DIAC business. It studied the official DIAC Twitter accounts in two ways. First, DIAC tweets are studied with text mining techniques to identify topics and their variations over time. And then, DIAC followers are analysed with social network analysis to find out how tweets spread over Twitter network. The methodology and techniques used in this work are general and can be easily applied to analysis of other Twitter accounts.

Keywords: Twitter, social media, text mining, topic modelling, social network analysis

1 Introduction

Twitter[1] is one of the most popular social media websites and has been growing rapidly since its creation in March 2006. As of March 2013, there were over 200 million active users, creating over 400 million tweets every day (Twitter Blog 2013). An advantage of Twitter is that it is real time and information can reach a large number of users in a very short time. As a result, there has been an increasing trend to analyse Twitter data in past years. One very early work on Twitter data analysis was published in 2007, which studied the topological and geographical properties of Twitter's social network and analysed user intentions at a community level (Java et al. 2007). It is followed by Kwak's work that analysed Twitter's follower-following topology, user ranking and top trending topics (Kwak et al. 2010). There were a lot of other publications on this topic recently (Bakshy et al. 2011, Poblete et al. 2011, Szomszor et al. 2011, Zubiaga et al. 2011, Bae & Lee 2012, Lehmann et al. 2012, Lu et al. 2012, Pennacchiotti et al. 2012, Stringhini et al. 2012, Tao et al. 2012, Chang et al. 2013).

However, there is little work reported on social media data analysis in government agencies. To analyse social media data and study the feasibility of using them to better understand and improve the business of the Australian Department of Immigration and Citizenship (DIAC)[2], a research project was performed recently in the Analytics Research Weeks in DIAC, with Twitter as a start point. This work studies the official DIAC Twitter accounts in two ways. At first, DIAC tweets are analysed with text mining techniques to identify topics and their variations over time. And then, DIAC followers are studied with social network analysis to find out how tweets spread over Twitter network. All analysis in this work was done with R[3] (R Core Team 2013) and several R packages.

The rest of this paper is organised as below. Section 2 introduces the Twitter data used in this work and also shows how to get data from Twitter. Tweets are then analysed with text mining and topic modelling in section 3. In section 4, DIAC followers are investigated with social network analysis and it demonstrates how tweets spread over Twitter network. Conclusions and discussions are provided in the last section.

2 Twitter Data

2.1 DIAC Twitter Accounts

There are two official Twitter accounts owned by DIAC, *@SandiHLogan* and *@DIACAustralia*. *@SandiHLogan* is an account of the DIAC National Communications Manager and used to be official account of DIAC. Its tweets were used in the analysis of text mining in section 3. In December 2012, a new dedicated account, *@DIACAustralia*, was created, and its data were used in social network analysis in section 4.

2.2 Getting Twitter Data

To pull data from Twitter, the *TwitterR* package[4] (Gentry 2013) was used, which provides an interface to the Twitter web API. In addition, the Twitter API v1's *"GET statuses/:id/retweeted_by/ids"*[5] was also used, together with the *RCurl* package (Lang 2013), to find out how tweets were retweeted. Because Twitter API v1 stopped in June 2013 and was replaced with v1.1, readers need to use *"GET statuses/retweeters/ids"* or *"GET statuses/retweets/:id"* provided in Twitter API v1.1[6] to trace the path on which a tweet was retweeted.

[1] http://www.twitter.com

[2] http://www.immi.gov.au

[3] http://www.r-project.org

[4] http://cran.r-project.org/web/packages/twitteR

[5] https://dev.twitter.com/docs/api/1

[6] https://dev.twitter.com/docs/api/1.1

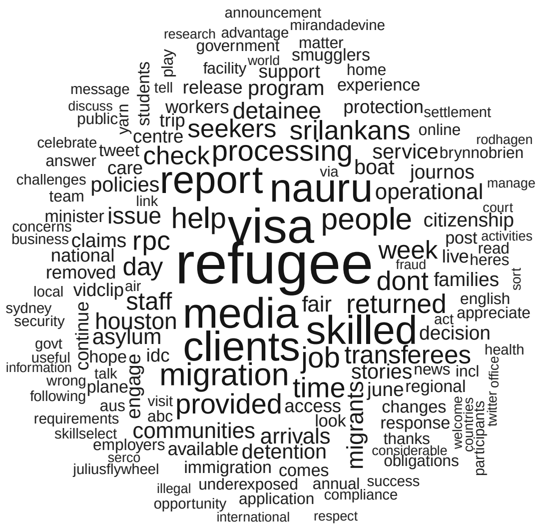

Figure 1: Word Cloud

3 Text Mining of Tweets

Tweets of *@SandiHLogan* extracted on 18 December 2012 were used in this analysis. At that time, it had over 6,000 Tweets and 5,493 followers, and followed 173 users. In this analysis, there were 1,409 tweets from 31 March to 9 December 2012, which were collected and analysed with text mining and topic modelling techniques to find topics and events over time. After that, topics and events were aligned with time series of visa applicants to find relationship between them.

The tweets were pulled from the Twitter website with R and the *twitteR* package (Gentry 2013), and then were processed and analysed with the *tm* package (Feinerer & Hornik 2013) and the *topicmodels* package (Grün & Hornik 2011), by following an example on text mining of Twitter data (Zhao 2013). At first, the text were cleaned by removing punctuations, numbers, hyperlinks and stop words, followed by stemming and stem completion. In addition to common English stop words, some other words, such as "Australia", "Aussie", and "DIAC", which appeared in most tweets, were also removed. After that, a term-document matrix was bulit and used for modelling. The results of text mining are shown in Figures 1, 2 and 3.

3.1 Frequent Terms and Term Network

Based on the term-document matrix, the frequency of terms was derived and plotted as a word cloud (see Figure 1), using the *wordcloud* package (Fellows 2013). In the word cloud, the more tweets a term appeared in, the bigger the term is shown. The figure shows that there were many tweets on refugee and skilled migration, and also some tweets on Sri Lankan and the Nauru Regional Process Centre (RPC).

Still based on the term-document matrix, a network of terms were built according to their co-occurrences in tweets, using the *Rgraphviz* package (Gentry et al. 2013). The result is shown in Figure 2. The vertices stand for terms, and the connections for the co-occurrences of terms in same tweets. A thick line indicates that the two corresponding terms appeared together in many tweets. The figure indicates that there are some tweets on transferees and the Nauru RPC, some on asylum seekers, some on a job fair in Houston, and some on a refugee week in June.

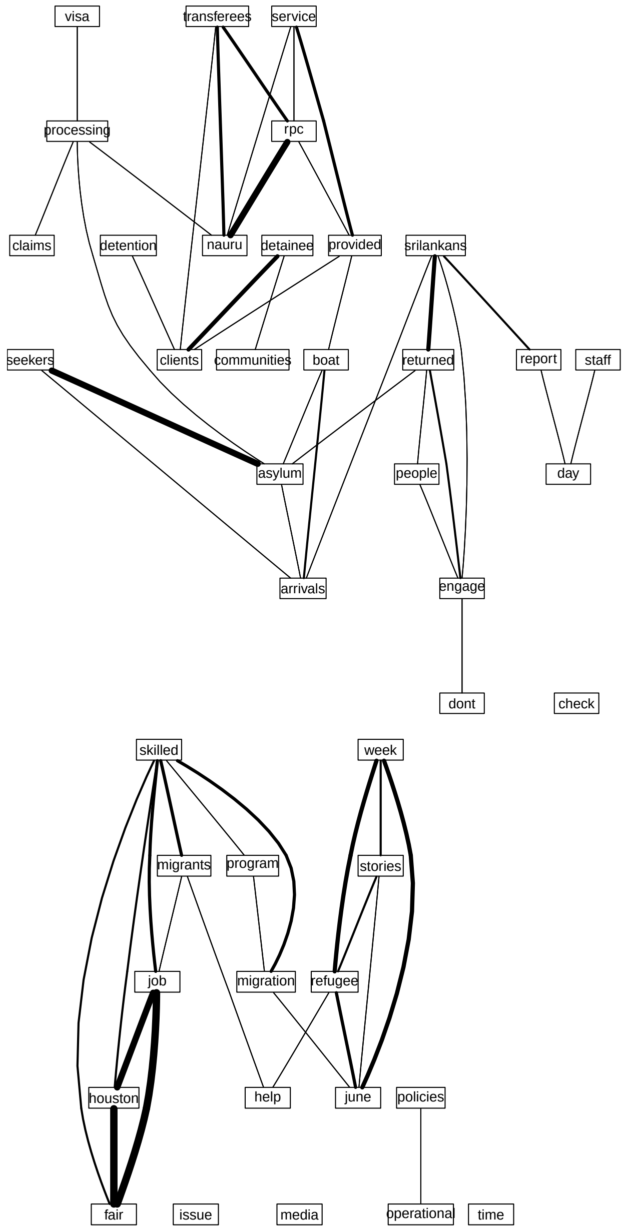

Figure 2: Term Network

3.2 Topic Modelling

After the above text mining of frequent terms and their connections, topics in tweets were studied. Topics were identified from tweets with the LDA (Latent Dirichlet Allocation) model (Blei et al. 2003) provided in the *topicmodels* package (Grün & Hornik 2011). Then the number of tweets in each topic was counted and plotted as a stream graph to show temporal variation of topics (see Figure 3).

Figure 3 can be taken as a stacked density plot of count of tweets on every topic, where the volume of tweets on a topic is shown with band width. Note that in the stream graph, the topics in legend are in the reverse order of those in the graph itself. That is, the first band stands for topic *staff, visa, media, changes* (the last one in legend), the second shows *skilled, job, fair, dont* (the 2nd last in legend), and so

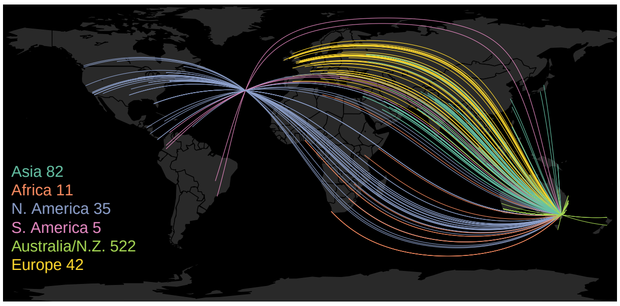

Figure 4: Twitter Follower Map

on. The figure shows that there were many tweets on refugee (see the 4th & 5th topics) in May and June 2012. In addition, the 2nd topic from the bottom in the stream graph shows many discussions on the Nauru RPC, transferee and Sri Lankan in November 2012.

A possible application is to align the stream graph with time series, such as the number of visa applications, to find out any relationship between them, and even further to predict the trend in the number of visa applications based on changes in topics and produce alerts for significant events and emerging topics.

4 Social Network Analysis of Twitter Followers and Retweeting

Following text mining of tweets in last section, this section studies DIAC Twitter account in the approach of social network analysis. This analysis focused on who the followers of *@DIACAustralia* were and how its tweets were retweeted by them and spread over the Twitter network. More specifically, its followers were investigated and shown on a word map, top retweeted messages were identified, and the spread of the above tweets and their potential impact was studied.

The Twitter data of *@DIACAustralia* and its followers were used in this social network analysis. This account started from December 2012, and data of it on 24 May 2013 were extracted. On that day, it had 118 Tweets and 1,493 followers and followed 91 users. The techniques involved are geomap, social network analysis and text mining, and the tools used are R, packages *twitteR* (Gentry 2013) and *igraph* (Csardi & Nepusz 2006), and the Twitter API. More details about how to extract Twitter data are provided in section 2.2.

4.1 Followers

Locations of followers were first checked. With the location information of Twitter accounts, a map of Twitter followers (see Figure 4) was produced using a *twitterMap* function[7] authored by Jeff Leek. The lines in the figures show the connections between DIAC (in Canberra) and its followers. Note that it shows only followers who have provided locations in their Twitter account descriptions.

Next, followers were categorised based on descriptions provided in their Twitter account, which give a short piece of information about owner of the account. Alternative ways to categorise followers are categorising based on their tweets, followers or friends (i.e., users that they follow). With text mining again, a term-document matrix was built for user descriptions of followers and then plotted as a term network (see

[7] http://biostat.jhsph.edu/~jleek/code/twitterMap.R

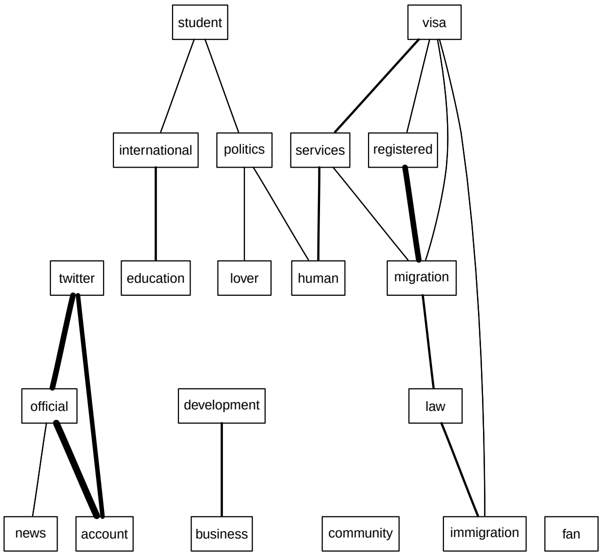

Figure 5: Term Network of Follower Descriptions

Figure 5). The figure shows some categories of followers. The subgraph on the bottom-left corner indicates some followers are official Twitter account of government departments or organisations. The top-left part shows that there are some followers who focus on international students and education. Another group of followers, shown in the right part of the figure, are registered migration agents, who provide visa and legal services.

After that, active and influential users among DIAC followers were inspected. Information of all followers of *@DIACAustralia* were collected, including when the accounts were created, how many tweets they had, how many users (i.e., followers) followed them, and how many users (i.e., friends) they were following. For every follower, the ratio of number of followers to number of friends was calculated, because an influential user tends to have a lot of followers but does not follow too many users. The average number of tweets of every follower per day is also calculated, which shows how active a user is. Based on the above numbers, a scatter plot of top followers were produced as Figure 6. Note that personal names are anonymised for privacy reasons. The figure shows that, the most influential and active followers largely fall into two categories.

- Media and correspondents: *7News Yahoo!7*, *The Morning Show* (on Channel 7) and *lia281* (an ABC Papua New Guinea Correspondent); and
- Government agencies and officials: *lat250* (quarrelling quandaries of question time, Parliament House Canberra), *ACPET* (national industry association for education and training), *AEC* (Australia Electoral Commission), *DFAT* (Department of Foreign Affairs and Trade, Australia), *Australian Customs*, *Sandi Logan* (DIAC National Communications Manager), *Aus745* (Australian Ambassador to US), etc.

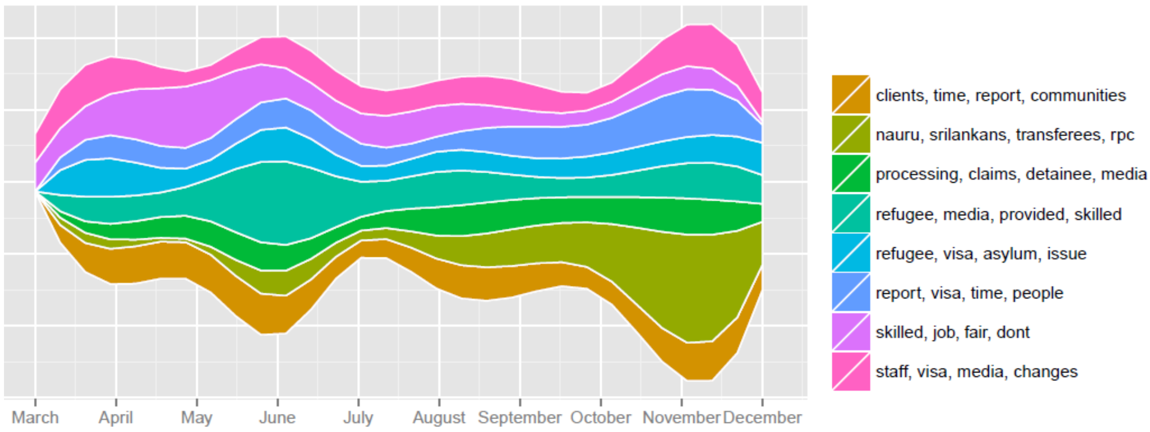

Figure 3: Stream Graph of Topics

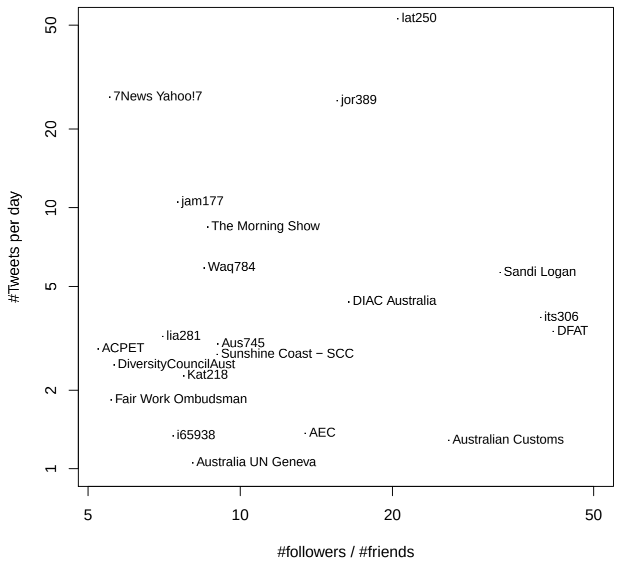

Figure 6: Top Influential and Active Followers

4.2 Tweets Most Retweeted and Their Spread Over Twitter Network

After studying who and where the followers were, this analysis presents what the most influential tweets were about and how they were retweeted via the Twitter network.

Figure 7 shows tweets that have been retweeted more than 10 times. The horizontal axis stands for time and the vertical for the number of times that a tweet was retweeted. The most retweeted one was a tweet on 9 January 2013 about Australia's low unemployment rate of migrants, and it was retweeted over 40 times.

The second most retweeted message was tweeted on 11 January 2013: "Are you an international student in Australia? New post-study work arrangements are being introduced http://t.co/g8c4yPlT". The URL in the tweet links to a post on the DIAC Migration Blog[8]. This tweet was investigated further to find out how it spread on Twitter network and how many users it reached. The retweeting data were extracted with the Twitter API mentioned in section 2.2, and the analysis were preformed with the *igraph* package (Csardi & Nepusz 2006).

Figures 8 and 9 show how the message were retweeted by DIAC followers and spread over Twitter network. Similar to Figure 6, personal names are anonymised and moreover, personal photos are replaced with an egg icon. Figure 8 shows a network of Twitter users who have retweeted the above message. Based on Figure 8, Figure 9 shows followers of those users and illustrates how many users the message might have reached. The message was retweeted by *DFAT*, who had 14,662 followers at that time, and then retweeted again by its followers, such as *deb338* (an Editor of ABC News Melbourne, 396 followers), *Austraining Int.* (352 followers) and *mym278* (a Policy Analyst at Chamber of Commerce & Industry, Queensland, 344 followers). The message was also retweeted by *Australia in UK* (Australian High Commission in UK, 1,129 followers) and then by *Dub706* (Australian Ambassador to Indonesia, 3,586 followers), who passed it on to his followers. In addition, it was also retweeted by other immediate followers of *@DIACAustralia*, such as *Ohj787* and *Sma346*, who were Immigration Officers at universities and education organisations. The above analysis shows that the message has potentially reached over 23,000 Twitter users.

5 Conclusions and Future Work

This paper presents a preliminary research on analysing DIAC Twitter data in approaches of text mining and social network analysis. With text mining of tweets, topics and their variations over time have been identified. Twitter followers have been analysed and the spread of tweets over Twitter network has been studied. With some initial interesting results identified, this research will be further studied in future research projects.

The methodology used in this work is general, the tools used are open-source software, and the data used in this work are publicly available on Twitter. Therefore, readers can easily replicate the analysis and apply it to Twitter accounts that they are interested in.

This research can be extended by analysing text from more Twitter accounts, analysing social network between them and their followers, and developing an effective method to find relationship between topics/events and variations in time series, e.g., the number of visa applicants, approvals, arrivals, departures and visa processing time.

It can also be extended further to investigate how messages spread, estimate their impacts and generate alerts. It would also be interesting to analyse tweets

[8] http://migrationblog.immi.gov.au

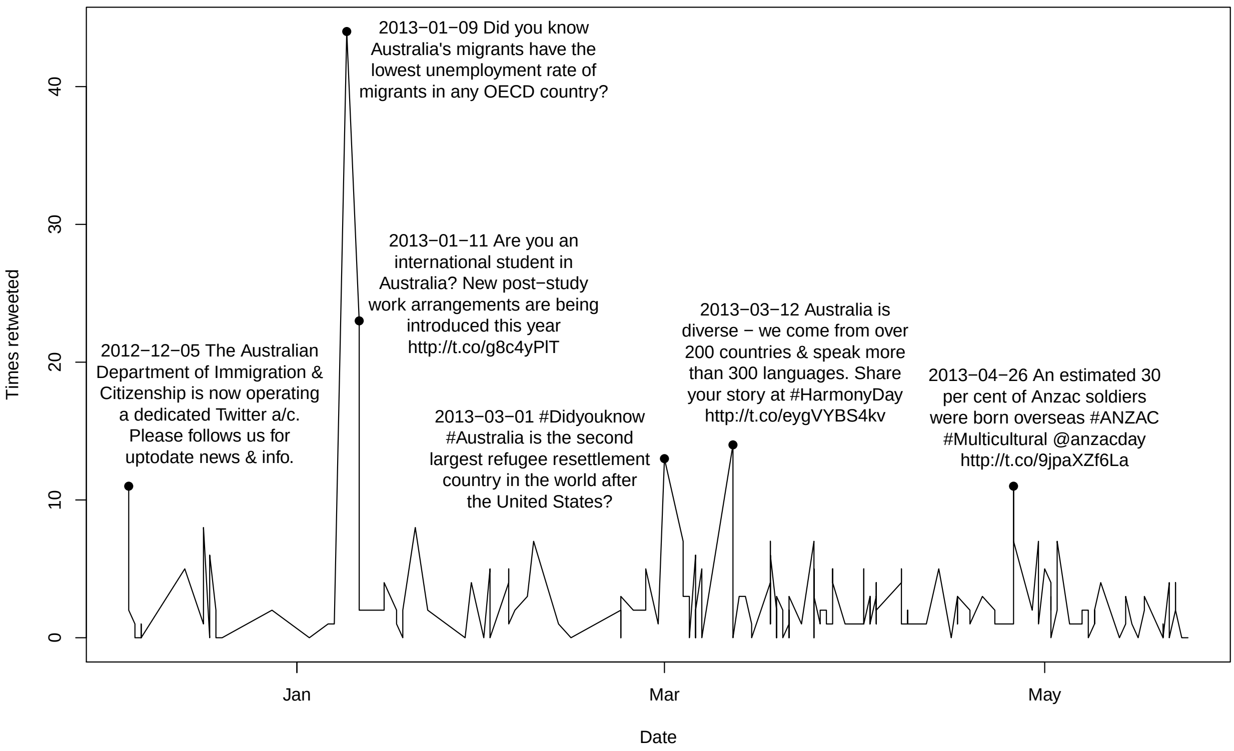

Figure 7: Tweets Most Retweeted

on specific topics based on Twitter hashtags, such as "#AustralianVisa" and "#refugee", and preform sentiment analysis for new legislations and policies.

Another possible future work is to study social network with data from multiple social media platforms, such as Twitter, Facebook[9] and Google+[10], and investigate interactions between government agencies, migration agencies and individuals.

Acknowledgements

I'd like to thank the Intent Management and Analytics Section, Department of Immigration and Citizenship for providing an opportunity to do this work in research weeks. I'd also like to thank Greg Hood from the Department of Agriculture, Fisheries and Forestry, and Fraser Tully and John D'arcy from the Department of Immigration and Citizenship for sharing their code for graph plotting.

References

Bae, Y. & Lee, H. (2012), 'Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers', *J. Am. Soc. Inf. Sci. Technol.* **63**(12), 2521–2535.
URL: *http://dx.doi.org/10.1002/asi.22768*

Bakshy, E., Hofman, J. M., Mason, W. A. & Watts, D. J. (2011), Everyone's an influencer: quantifying influence on twitter, *in* 'Proceedings of the fourth ACM international conference on Web search and data mining', WSDM '11, ACM, New York, NY, USA, pp. 65–74.
URL: *http://doi.acm.org/10.1145/1935826.1935845*

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of Machine Learning Research* **3**, 993–1022.
URL: *http://dl.acm.org/citation.cfm?id=944919.944937*

Chang, Y., Wang, X., Mei, Q. & Liu, Y. (2013), Towards twitter context summarization with user influence models, *in* 'Proceedings of the sixth ACM international conference on Web search and data mining', WSDM '13, ACM, New York, NY, USA, pp. 527–536.
URL: *http://doi.acm.org/10.1145/2433396.2433464*

Csardi, G. & Nepusz, T. (2006), 'The igraph software package for complex network research', *InterJournal* **Complex Systems**, 1695.
URL: *http://igraph.sf.net*

Feinerer, I. & Hornik, K. (2013), *tm: Text Mining Package*. R package version 0.5-8.3.
URL: *http://CRAN.R-project.org/package=tm*

Fellows, I. (2013), *wordcloud: Word Clouds*. R package version 2.4.
URL: *http://CRAN.R-project.org/package=wordcloud*

Gentry, J. (2013), *twitteR: R based Twitter client*. R package version 1.1.6.
URL: *http://CRAN.R-project.org/package=twitteR*

Gentry, J., Long, L., Gentleman, R., Falcon, S., Hahne, F., Sarkar, D. & Hansen, K. D. (2013), *Rgraphviz: Provides plotting capabilities for R graph objects*. R package version 2.4.1.

[9] http://www.facebook.com
[10] http://plus.google.com

Grün, B. & Hornik, K. (2011), 'topicmodels: An R package for fitting topic models', *Journal of Statistical Software* **40**(13), 1–30.
URL: *http://www.jstatsoft.org/v40/i13/*

Java, A., Song, X., Finin, T. & Tseng, B. (2007), Why we twitter: understanding microblogging usage and communities, *in* 'Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis', WebKDD/SNA-KDD '07, ACM, New York, NY, USA, pp. 56–65.
URL: *http://doi.acm.org/10.1145/1348549.1348556*

Kwak, H., Lee, C., Park, H. & Moon, S. (2010), What is Twitter, a social network or a news media?, *in* 'WWW '10: Proceedings of the 19th international conference on World wide web', ACM, New York, NY, USA, pp. 591–600.

Lang, D. T. (2013), *RCurl: General network (HTTP/FTP/...) client interface for R*. R package version 1.95-4.1.
URL: *http://CRAN.R-project.org/package=RCurl*

Lehmann, J., Gonçalves, B., Ramasco, J. J. & Cattuto, C. (2012), Dynamical classes of collective attention in twitter, *in* 'Proceedings of the 21st international conference on World Wide Web', WWW '12, ACM, New York, NY, USA, pp. 251–260.
URL: *http://doi.acm.org/10.1145/2187836.2187871*

Lu, R., Xu, Z., Zhang, Y. & Yang, Q. (2012), Life activity modeling of news event on twitter using energy function, *in* 'Proceedings of the 16th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part II', PAKDD'12, Springer-Verlag, Berlin, Heidelberg, pp. 73–84.
URL: *http://dx.doi.org/10.1007/978-3-642-30220-6*$_7$

Pennacchiotti, M., Silvestri, F., Vahabi, H. & Venturini, R. (2012), Making your interests follow you on twitter, *in* 'Proceedings of the 21st ACM international conference on Information and knowledge management', CIKM '12, ACM, New York, NY, USA, pp. 165–174.
URL: *http://doi.acm.org/10.1145/2396761.2396786*

Poblete, B., Garcia, R., Mendoza, M. & Jaimes, A. (2011), Do all birds tweet the same?: characterizing twitter around the world, *in* 'Proceedings of the 20th ACM international conference on Information and knowledge management', CIKM '11, ACM, New York, NY, USA, pp. 1025–1030.
URL: *http://doi.acm.org/10.1145/2063576.2063724*

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Stringhini, G., Egele, M., Kruegel, C. & Vigna, G. (2012), Poultry markets: on the underground economy of twitter followers, *in* 'Proceedings of the 2012 ACM workshop on Workshop on online social networks', WOSN '12, ACM, New York, NY, USA, pp. 1–6.
URL: *http://doi.acm.org/10.1145/2342549.2342551*

Szomszor, M., Kostkova, P. & Louis, C. S. (2011), Twitter informatics: Tracking and understanding public reaction during the 2009 swine flu pandemic, *in* 'Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01', WI-IAT '11, IEEE Computer Society, Washington, DC, USA, pp. 320–323.
URL: *http://dx.doi.org/10.1109/WI-IAT.2011.311*

Tao, K., Abel, F., Gao, Q. & Houben, G.-J. (2012), Tums: twitter-based user modeling service, *in* 'Proceedings of the 8th international conference on The Semantic Web', ESWC'11, Springer-Verlag, Berlin, Heidelberg, pp. 269–283.
URL: *http://dx.doi.org/10.1007/978-3-642-25953-1*$_2$2

Twitter Blog (2013), 'Celebrating #twitter7'.
URL: *https://blog.twitter.com/2013/celebrating-twitter7*

Zhao, Y. (2013), 'Using text mining to find out what @RDataMining tweets are about'.
URL: *http://www.rdatamining.com/examples/text-mining*

Zubiaga, A., Spina, D., Fresno, V. & Martínez, R. (2011), Classifying trending topics: a typology of conversation triggers on twitter, *in* 'Proceedings of the 20th ACM international conference on Information and knowledge management', CIKM '11, ACM, New York, NY, USA, pp. 2461–2464.
URL: *http://doi.acm.org/10.1145/2063576.2063992*

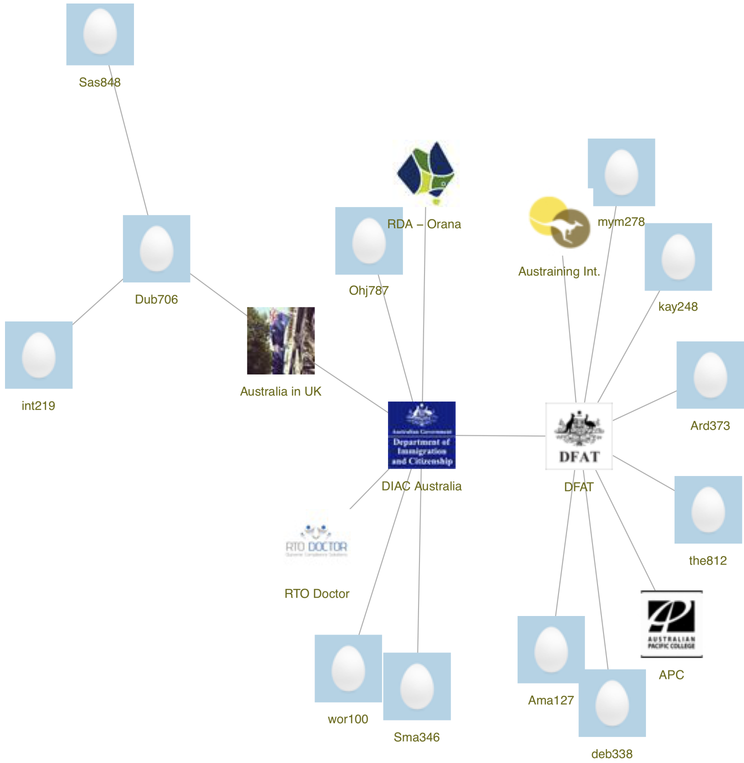

Figure 8: Retweet Graph - I

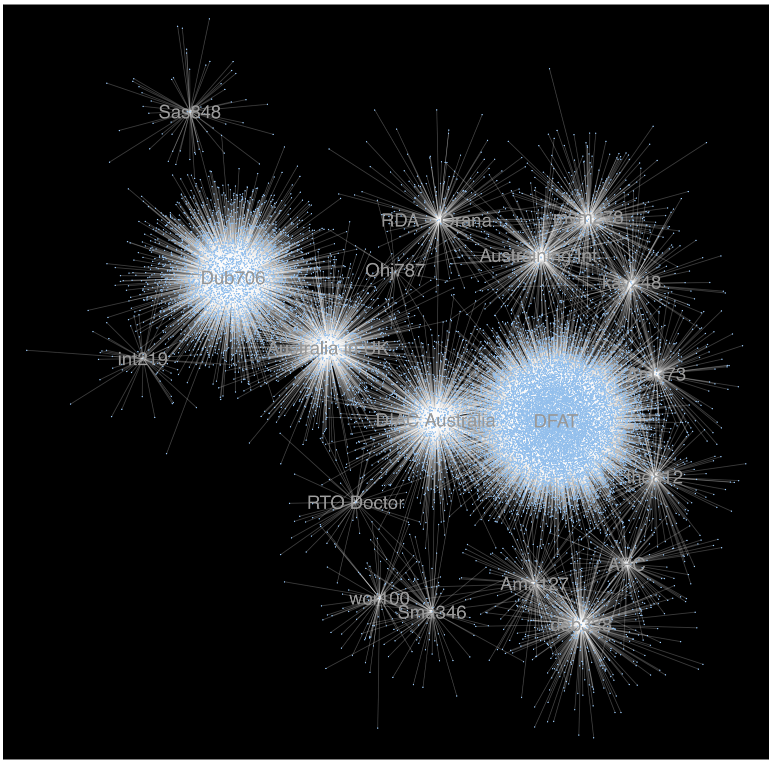

Figure 9: Retweet Graph - II