CHARACTERISING TWITTER REPLY CHAINS IN THE AUSTRALIAN TWITTERSPHERE

Brenda Moon
Queensland University of Technology

Introduction

This research contributes to the understanding of conversation on Twitter by looking at reply chain networks in the Australian Twittersphere. By examining all of the reply chains in November 2016 from the comprehensive TrISMA dataset this paper develops a description of the characteristics of reply chains in the Australian Twittersphere.

Many studies of conversation on Twitter have been based on searching for specific keywords or hashtags, potentially missing important sections of the conversation (Rambukkana, 2015; Bruns & Burgess, 2015). These keyword or hashtag based studies by their nature focus on a single topic and may not consider the conversation chains at all (Bruns 2012; Moon, Suzor & Matamoros Fernández 2016).

To address this considerable limitation, Lorentzen and Nolin (2015) proposed a method of collecting all the tweets that were replied to by the tweets collected using a keyword or hashtag using the Twitter API. By using the Tracking Infrastructure for Social Media Analysis (TrISMA) (Bruns, Burgess & Banks et al., 2016), Moon, Suzor & Matamoros Fernández (2016) extended this technique to recursively collect both tweets replied to and in reply to the initial set of keyword tweets. TrISMA captures the public tweets of 4m Australian accounts on a continuous basis, and allows searching both for tweets which are in reply to the seed tweets and for tweets which the seed tweets replied to. In comparison, the Twitter API only supports retrieving tweets which are replied to by the seed tweet.

The present paper both advances beyond the limitation to specific pre-identified topics as expressed through keywords and hashtags, and builds on the author's previous work in developing methods to trace conversations both backwards and forwards from selected seed tweets. This results in a considerably more complete picture of Twitter

conversations than has been developed in the past. However, one remaining limitation of the TrISMA dataset is that it only includes tweets sent by the accounts identified as Australian, and so conversation chains will break if there are tweets which are external to the Australian Twittersphere, even if they then continue inside the Australian Twittersphere.

Identifying tweets

Tweets involved in reply chains were first identified by selecting all the tweets from the TrISMA dataset that had their 'in_reply_to_status_id' field set, which contains the unique tweet ID for the replied-to tweet. For the purpose of this paper, we limited the timeframe to tweets posted in November 2016.

The reply chains are then identified by using the 'in_reply_to_status_id' field for each tweet to retrieve the replied-to tweet, if it is already in the replies dataset as shown in Figure 1. Each reply chain on Twitter has an original tweet (indicated in green in Figure 1) that is not a reply, i.e. it does not have the 'in_reply_to_status_id' field set.
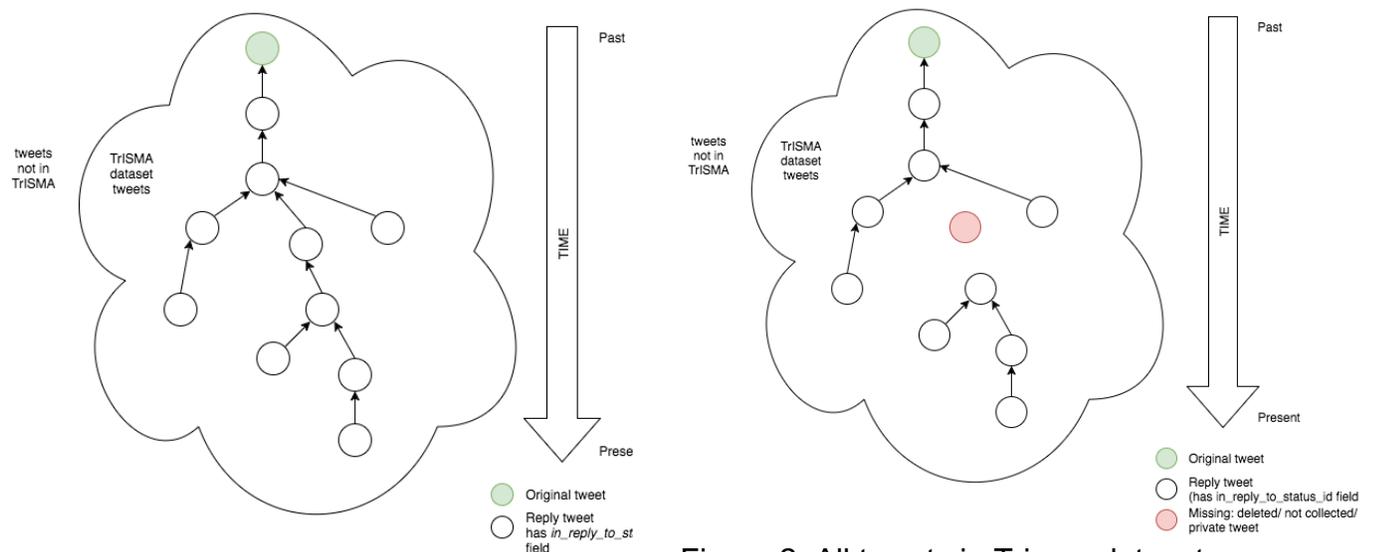

Figure 1. Reply chain with all tweets in TrISMA dataset

Figure 2. All tweets in Trisma dataset – missing tweet

Figure 2 shows the same reply chain, but with one tweet missing on TrISMA (shown in red). The missing tweet means that we find two reply chains, and have no way of knowing that they are connected.

Tweets which are replied to and tweets which are replies that are outside of the TrISMA dataset (shown in Figure 3 in blue and purple respectively) are excluded from this study.
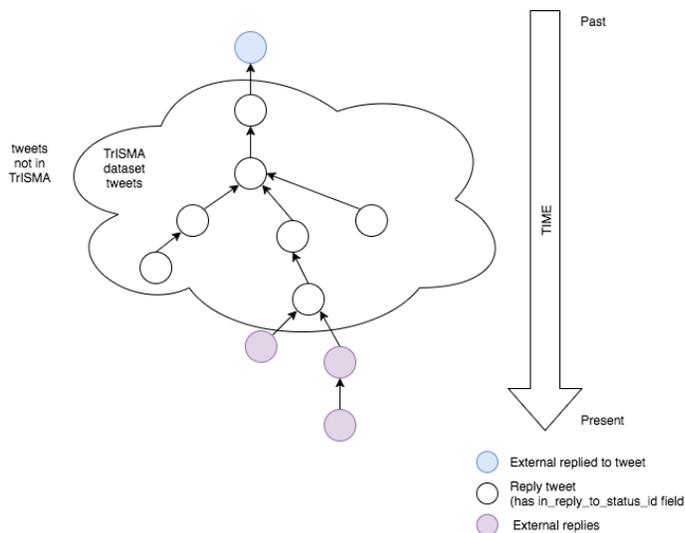
Figure 3. Tweets outside TrISMA dataset

Analysing reply chains

Reply chains features such as the number of tweets in the chain, duration of the conversation, the interval between tweets, branching structure, and number of participating accounts, as well as tweet content features such as links, topics, and images are examined to identify characteristics which can be used to classify the reply chains.

Initial Results

In November 2016, there are 39.4m tweets in the TrISMA dataset, of which 7.1m (18%) are replies with their 'in_reply_to_status_id' set.

6.2m unique tweet ids are listed in 'in_reply_to_status_id' fields during this time. Of these, 1.2m (20%) are included in the November 2016 'in_reply_to_status_id' dataset, and 0.9m (15%) are original Australian tweets posted during November 2016. An additional 21,500 (0.4%) replied-to tweets were found in the TrISMA dataset but were posted before November 2016; these were added to the overall dataset. 4.1m (65%) of the tweets that Australian accounts replied to during November 2016 originated from outside the TrISMA dataset, indicating that substantial amount of the conversation in the Australian Twittersphere also includes participation by non-Australian accounts.

I will present an analysis of these different metrics for reply chains in the Australian Twittersphere and evaluate their utility in categorising their different conversational patterns. This may make it possible to computationally distinguish single-authored threads, rapid and heated arguments, extended chats, multi-user responses to a single controversial tweet, and other types of conversation. This has important applications in large-scale, automated social media discourse analysis and in the computational filtering of 'big social data' into smaller samples for manual analysis.

Limitations

The initial results show that 65% of the replied-to tweets are not in the TrISMA dataset (shown in blue in Figure 3), which is limited to tweets from Australian accounts. Retrieving these tweets from the Twitter API would allow us to check which were original tweets and which were replies, and this could be done recursively to find the older tweets in the reply chain. External tweets that are replies (shown in purple in Figure 3) to the tweets in the reply chain can only be detected manually through the Twitter client interface, not in the TrISMA dataset or using the Twitter API. This study focusses on the Australian Twittersphere so we exclude these external tweets, even though that may reduce the lengths of some of the reply chains.

Because this study uses reply chains, it does not consider other forms of conversation on Twitter – such as manual replies, mentions, retweets, and quoted retweets. It would be possible to extend this study to quoted retweets and button retweets as these generate metadata in fields similar to the 'in_repy_to_status_id' field (including 'quoted_status_id' and 'retweeted_status', the latter of which contains the full retweeted tweet), and this is an area for further work.

Acknowledgments

References

Bruns, A. (2012). How long is a Tweet? Mapping Dynamic Conversation Networks on Twitter using Gawk and Gephi. *Information, Communication & Society*, *15*(9), 1323–1351. http://doi.org/10.1080/1369118X.2011.635214

Bruns, A., and Burgess, J. 2015. Twitter hashtags from ad hoc to calculated publics. In *Hashtag Publics: The Power and Politics of Discursive Networks*, N. Rambukkana, Ed.New York: Peter Lang, 13-28

Bruns, A., Burgess, J., Banks, J., Tjondronegoro, D., Dreiling, A., Hartley, J., Leaver, T., Aly, A., Highfield, T., Wilken, R., Rennie, E., Lusher, D., Allen, M., Marshall, D., Demetrious, K., and Sadkowsky, T. (2016). *TrISMA: Tracking Infrastructure for Social Media Analysis*. Retrieved from http://trisma.org/

Moon, B., Suzor, N., Matamoros-Fernandez, A. (2016, October 5-8). Beyond Hashtags: Collecting and analysing conversations on Twitter. Paper presented at AoIR 2016: The 17th Annual Meeting of the Association of Internet Researchers. Berlin, Germany: AoIR. Retrieved from http://spir.aoir.org.

Lorentzen, D. G., & Nolin, J. (2015). Approaching Completeness: Capturing a Hashtagged Twitter Conversation and Its Follow-On Conversation. *Social Science Computer Review*. http://doi.org/10.1177/0894439315607018

Rambukkana, N. (2015). *Hashtag Publics: The Power and Politics of Discursive Networks*. Peter Lang Publishing, Incorporated.