Conversations on Twitter: Structure, Pace, Balance

Danica Vukadinović Greetham[1] and Jonathan A. Ward[2]

[1] Centre for the Mathematics of Human Behaviour
Department of Mathematics and Statistics
University of Reading, UK
d.v.greetham@reading.ac.uk
[2] Department of Applied Mathematics
University of Leeds, UK
j.a.ward@leeds.ac.uk

Abstract. Twitter is both a micro-blogging service and a platform for public conversation. Direct conversation is facilitated in Twitter through the use of @'s (mentions) and replies. While the conversational element of Twitter is of particular interest to the marketing sector, relatively few data-mining studies have focused on this area. We analyse conversations associated with reciprocated mentions that take place in a data-set consisting of approximately 4 million tweets collected over a period of 28 days that contain at least one mention. We ignore tweet content and instead use the mention network structure and its dynamical properties to identify and characterise Twitter conversations between pairs of users and within larger groups. We consider conversational balance, meaning the fraction of content contributed by each party. The goal of this work is to draw out some of the mechanisms driving conversation in Twitter, with the potential aim of developing conversational models.

Keywords: Twitter mentions networks, conversations models, maximal cliques

1 Introduction

The rapid uptake of online social media, combined with consumer behavioural changes around television and news broadcasting, has instigated a sea change in attitudes within the advertising and marketing sectors. A frequently encountered adage is that "everything is about conversation and not about broadcasting" [10,6]. By facilitating public addressability through the @ sign (so called 'mentions') and enabling private messages, Twitter has confirmed their intention to function as a communication channel as well as a broadcasting tool. Access to large quantities of data produced by Twitter users has resulted in a surge of interest from the academic community [20], who have largely focused on Twitter's information flow and retweet behaviour, and hence implicitly the underlying network of 'followers' (e.g. [22,21]). While broadcasting short messages, or micro-blogging, remains an important component of Twitter use, to

our knowledge comparatively little work has addressed the mining of (public) conversations on a large scale [3,19,14]. Consequently, we focus in this paper on analysing the network of communication patterns resulting from mentions in Twitter.

Although it may not always be clear, even from message content, what intention a user had in mind when posting—information seeking or information sharing, broadcasting or conversation—we have tried to specifically extract conversations by focusing our data-analysis on reciprocated tweets. Moreover, we have completely ignored the content of conversations and concentrated on structural and dynamic properties of the underlying mentions network. Our main objective was to mine actionable insights that could inform our knowledge of conversational mechanisms and the frequency/timings of tweets. Our hope is that empirical observations and quantifiable insights from this analysis could inform a simple, data driven model of the timing and structure of Twitter conversations. One possible application would be for automated recommendations of conversation trends, as discussed in [3,1].

A large number of registered Twitter accounts are operated by automated software scripts, known as *bots* [18]. While such accounts are encouraged for the purpose of developing applications and services, bots whose functions violate Twitter policy (e.g. spammers) are common. The analysis of conversational patterns and the development of associated models have potential application for those trying to develop algorithms that can identify nuisance bots. Furthermore, the identification of groups of Twitter users who, through conversational behaviour, are particularly influential on a specific topic would be particularly attractive in the marketing sector. Thus, understanding conversational structure could impact the design and implementation of social media campaigns and potentially provide a quantitative comparison between Twitter discourse and other channels of communication, such as face-to-face, telephone, SMS, forums or email. In addition, curating and recommending conversational trends, for both Twitter and more generally in online social media, is crucial for social networking sites as it is one of the main characteristics of user experience. We believe that a better understanding of the structure, dynamics and balance of multi-user conversation is key to improving such automated curation systems. Ultimately, we hope that studying Twitter conversation can ultimately improve user experience.

In Section 2, we give an account of previous work in this space. Our results of pairwise and multiple conversations and the Twitter dataset we used are presented in Section 3. Finally, in Section 4 we summarise and describe possible directions of future work.

2 Previous work

The phenomenal uptake of Twitter over the last few years has resulted in a rapidly growing interest in mining Twitter data and particularly sentiment analysis of tweets. A recent study analyzing a large amount of Twitter and Face-

book data [12] found correlations between friendship/follower relations and positive/negative moods of Twitter users. Diurnal and seasonal mood rhythms that are common across different cultures have also been identified in cross-cultural Twitter data [5], shedding light on the dynamics of positive and negative affect.

A study of conversations within a sample of 8.5k tweets collected over an hour long period [9] found that the @ sign appeared in about 30% of the collected sample, its function was mostly for addressing (as intended) and it was relatively well reciprocated—around 30% of messages containing an @ were reciprocated within an hour. The majority of these conversations were short, coherent exchanges between two people, but longer exchanges did occur, sometimes consisting of up to 10 people. They found that

> *"...Tweets with @ signs are more focused on an addressee, more likely to provide information for others, and more likely to exhort others to do something—in short, their content is more interactive. "*

Twitter conversations also contain both momentarily salient or 'peaky' topics, signified by increased word-use frequency of specific terms, as well as more 'persistent conversations', in which less salient terms recur over longer periods [14]. In addition, words that relate to negative emotions are less persistent [22].

In [3], several algorithms for recommending conversations based on the lengths, topic and 'tie-strength'[3] of conversations were compared. Their results showed that the different uses of Twitter (social vs. informational) had a big influence on the algorithm's performance — recommendations based on tie strength were preferred by social users, whilst those based on topic were preferred by informational users. Related work considered automated curation of online conversations to present discussion threads of interest to users in e.g. Facebook and Google+. [1]. Key to this was the prediction of conversation length around a topic and re-entry of interlocutors. In another work concerning Twitter conversation [13], a relatively large corpus and content (topic) analysis of 1.3 million tweets was used to develop an unsupervised model of dialogue from open-topic data.

In our work we completely ignore content, instead focusing on timing, structure and balance of conversation between pairs of individuals as well as multi-user conversations. Our contribution is an attempt to map the structure of Twitter exchanges over a relatively large dataset, while offering some new methods to mine conversation data and improve statistical models of dialogue.

3 Analysis

3.1 Data

The Twitter data-set investigated in this paper was collected on our behalf by Datasift, a certified Twitter partner, allowing us to access the full Twitter

[3] Tie-strength is an increasing function of the number of exchanged messages between two people and the number of messages exchanged between them and their mutual friends.

firehose rather than being rate-limited by the API. The data-set consists of all UK based[4] Twitter users that sent tweets with at least one mention between 8 Dec 2011 and 4 Jan 2012 (28 days in total). In the remainder of the paper, use of the word 'tweet' will specifically mean tweets containing at least one mention. Mentions are messages that include an @ followed by a username. Thus if person a puts "@b", it designates that a is addressing the tweet to b specifically. Mentions are not private messages and can be read by anyone who searches for them. A tweet can be addressed to several users simultaneously using @ repetitively. Any Twitter user can mention any other Twitter user, they don't have to be related in any way. Since conversational characteristics are influenced by many factors, including language, culture, community membership etc., one has to keep in mind the natural limitations of the results of our analysis.

We preprocessed the data, removing empty mentions and self-addressing[5] and created a directed multigraph, or mentions network, containing $3,614,705$ timestamped arcs (individual mentions) from a total of $819,081$ distinct usernames, or nodes. Of these distinct usernames, $732,043$ were "receivers", i.e. to whom a message was addressed, and $137,184$ were "tweeters", i.e. people who tweeted a message with a mention. There were approximately 50k nodes that appeared both as tweeters and receivers. Note that our graph is a multigraph, meaning that multiple arcs are allowed between pairs of nodes, each having a direction and timestamp.

3.2 Conversations

An important feature of both face-to-face conversation [16,15] and computer-mediated communication [8], is the process of turn-taking. Thus in sequences of mentions between pairs of users, say a and b, we might expect that sequences like $ABABAB$ would be more common than say $AAABBB$, where we use A to denote that party a mentions party b and likewise B to denote that party b mentions party a.

To establish if this is the case, we assume the null hypotheses that contributions are independent events with probability P_A that party a contributes to a conversation and thus probability $P_B = 1 - P_A$ that party b contributes. For a given interaction sequence of length N between parties a and b, we are interested in the number of occurrences of B following A and vice-versa. We call these *transitions*, thus the sequence $ABAABBA$ of length $N = 7$, has 4 transitions. Note that we focus on reciprocated interactions, meaning that each party makes at least one contribution and consequently that there is by default at least one transition in all interactions that we consider. We call the remaining transitions the *excess transitions*. For any sequence of length N, the maximum possible number of excess transitions is clearly $N - 2$. Under the null hypotheses, excess

[4] All Twitter users appearing in our data-set had selected the UK as their location.

[5] Self-mentioning was surprisingly common in the data-set: 12,680 different users created a total of 44,319 self-mentions, with the maximum being 5,586 from an automated service that advertises itself at the end of each tweet.

transitions occur with probability $P_T = 2P_A(1 - P_A)$. Since we assume that transitions are independent, the probability distribution of a given number of excess transitions is binomial, and thus the expected number is $E_T = (N-2)P_T$ with variance $V_T = (N-2)P_T(1-P_T)$.

To test the null hypothesis, we consider all reciprocated pairwise interaction sequences in our Twitter data-set. For each sequence having n_X contributions from party $X \in \{A, B\}$, we assume that the probability of party a contributing is simply $n_A/(n_A + n_B)$. This does not yield any problematic probabilities (i.e. 0 or 1) since both parties always make at least one contribution.

Each sequence may have a different number of interactions and a different transition probability, but assuming that the pairwise interactions are independent, the expectation and variance of the ensemble is simply equal to the sum of the interaction expectations and variances respectively. Doing this, we find that the expected number of transitions is 85,390 with a standard deviation of 226.3, but we observe 88,758 transitions in practice, more than 15 standard deviations above the expected value. We take this as strong evidence that we can reject the null hypothesis and thus infer that the data contains a significant level of turn-taking and hence conversation.

Each sequence of pairwise interactions may constitute a number of different conversations, but ascertaining when one conversation ends and another begins may be an extremely difficult task, especially when the goal is to apply an automated processes to a large data-set. Instead of using a time-intensive lexical analysis, we investigate whether we can detect conversations by applying a simple threshold rule to the time gap between responses, where we assume that a time gap that is larger than the threshold indicates the start of a new conversation.

This method requires that we can identify a suitable threshold. To achieve this, we divide each sequence of pairwise interactions up according to a given threshold, then define distinct conversations to be reciprocated sub-sequences, i.e. sequences containing a contribution from both parties. Thus the number of sub-sequences n_{I} is always larger than the number of distinct conversations n_{C}. In Fig. 1(a) and (b) we plot the mean number of sub-sequences and the mean number of distinct conversations respectively over a range of threshold values. The number of distinct conversations n_{C} has a peak value at approximately 9hrs. This peak is expected, since we only count reciprocated interactions as distinct conversations. Thus small threshold values, which split an interaction sequence up into a large number of short sub-sequences (see Fig. 1(a)), result in relatively few distinct conversations because many of the sub-sequences feature contributions from only one party. High threshold values also result in a small number of conversations, but this is simply because they do not split the sequence up into many sub-sequences. Thus the maximum at 9hrs is a natural choice of threshold and corresponds to one's intuition that conversations may reflect diurnal patterns.

The mean and median number of tweets during conversations were 13.09 and 4 respectively, but the distribution was heavy tailed (see Fig. 2).

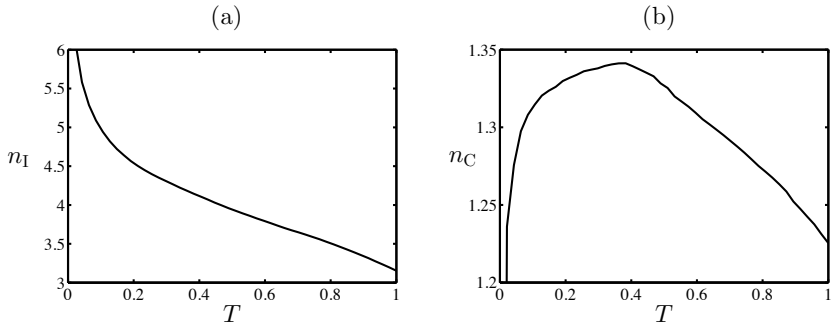

Fig. 1: Panel (a): Mean number of subsequences for a range of threshold values. Panel(b): Mean number of distinct conversations for a range of threshold values. Note that T, time threshold in hours, is normalised on the x-axis.

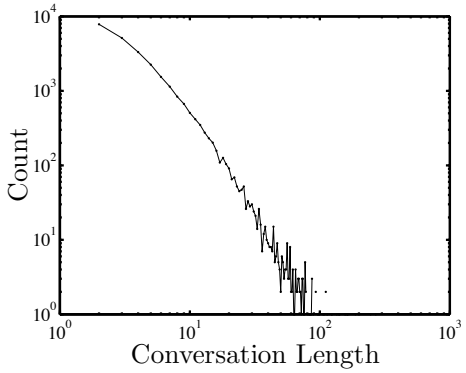

Fig. 2: Distribution of conversation length.

We now consider whether the number of contributions from each party are similar, or 'balanced' within pairwise interactions and conversations. For a given sequence of tweets, there are two ways to compute balance, we can either consider the ratio of means $b = \langle \max(n_A, n_B) \rangle / \langle \min(n_A, n_B) \rangle$ or the mean of ratios $\beta = \langle \max(n_A, n_B) / \min(n_A, n_B) \rangle$. We will use the subscripts 'I' and 'C' to denote whether these have been calculated for interactions or conversations respectively. Since we only consider reciprocated interactions, both quantities are well-defined and we would generally expect $b < \beta$. For the total number of interactions between pairs, we find that $b_{\mathrm{I}} = 2.424$ and $\beta_{\mathrm{I}} = 3.457$. Thus on average, one party contributes around 3 times as much as the other. For the sub-set of conversations, we find that $b_{\mathrm{C}} = 1.148$ and $\beta_{\mathrm{C}} = 1.425$. These are much closer to 1, and hence more what we would expect from typical, balanced conversations. The distribution of conversation contribution ratios is plotted in Fig. 3(a), which illustrates that conversations are most likely to be balanced, but some extremely unbalanced conversations do occur. In Fig. 3(b), for each

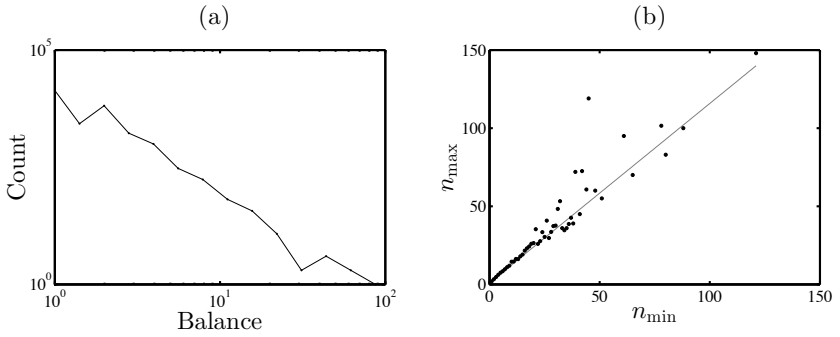

Fig. 3: Panel (a): Distribution of conversation balance. Panel (b): Mean maximum conversation contribution as a function of minimum contribution.

minimum conversation contribution $n_{\min} = 1, 2, 3, \ldots$, we compute the mean of the maximum contribution $n_{\max}$. There is a roughly linear trend (the grey line is $n_{\max} = 1.148 n_{\min} + 1$), which further illustrates conversational balance.

3.3 Multi-user conversations

By allowing multiple @ signs in one message, a Twitter user could send a tweet to several recipients simultaneously, facilitating multi-user conversations or *multicasting*. Note that because of the 140 character limit there is a physical limit on how many users each message can be multicast to.

In this part of analysis, our aim is to

- Identify multi-users exchanges;
- Determine how many users typically engage in them;
- Identify their time-frame, pace and how balanced they are.

In addition, are all users equally involved, or do some dominate the discussion? Are the same people at the heart of different multi-user conversations? What are the enablers and inhibitors of conversation flowing in the sense of pauses between consecutive contributions?

3.4 Identification of multi-users conversations

The reciprocated mentions data represents a directed multi-graph G (where an edge from A to B implies at least one edge from B to A), thus multi-user exchanges correspond to strongly-connected[6] subgraphs of G with $k > 2$ participants. We ran a non-recursive version of Tarjan's algorithm [17,11], as

[6] A directed graph is called strongly-connected if there is a path from each vertex in the graph to every other vertex. This means that for two vertices a and b there is a path in both directions, i.e. from a to b and also from b to a. Strongly-connected components of a graph are maximal subgraphs that are strongly-connected.

implemented in NetworkX [7], to get a list of the strongly-connected components of G. Pairwise conversations were discussed in Section 3.2, so we excluded all strongly-connected components of size 2 from the present analysis. Each strongly-connected component of at least three vertices was then transformed into an *undirected* multi-graph and we ran the NetworkX implementation of the modified Bron's algorithm [2] to find all maximal cliques[7]. We then disregarded all cliques of size two. We found in total 2190 cliques of size 3, 4, 5 and 6. The total number of users in these cliques was 3275 which is around 20% of users who reciprocated mentions.

In order to take the time elapsed between consecutive messages into account, we use the same threshold method explained in subsection 3.2, this time demanding for an exchange to be a"conversation" that there is a contribution from all parties and got relatively similar results (see Fig 4). The number of exchanges which had contribution from all parties was at peak around 9 and 11 hours. We took a threshold of 9 hours which gave us 334 multiuser conversations of sizes 3, 4, and 5 (see Fig 5a).

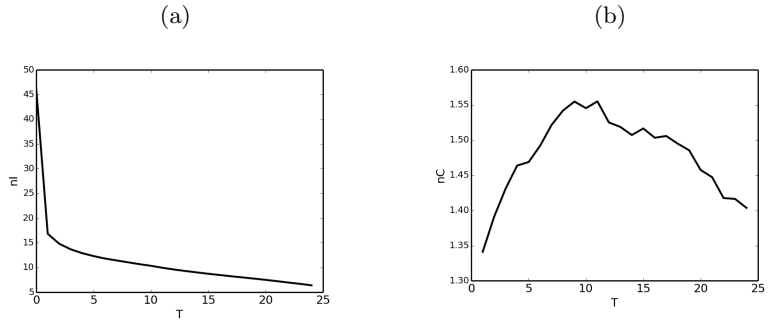

Fig. 4: Panel (a): Mean number of subsequences for a range of threshold values. Panel (b): Mean number of distinct conversations for a range of threshold values (threshold T in hours).

Most users (out of 646) in our dataset were involved in just one multi-user conversation, but a small number were involved in multiple conversations. The users' involvement in multi-user conversation is illustrated in Fig 5b.

When examining the time-frame of multi-user exchanges, we found that the correlation coefficient between the total number of exchanges between clique members and the average difference between consecutive exchanges was -0.244 (see Fig 6a). This was not surprising, since we would expect lively conversations (with lots of exchanged messages) to have a relatively fast pace, in contrast to a casual exchange of messages with longer differences inside our chosen 9 hour time-window. The same picture is obtained from looking at the median time differences between consecutive messages across different clique sizes (see Fig

[7] Maximal cliques are the largest complete subgraphs containing a given node.

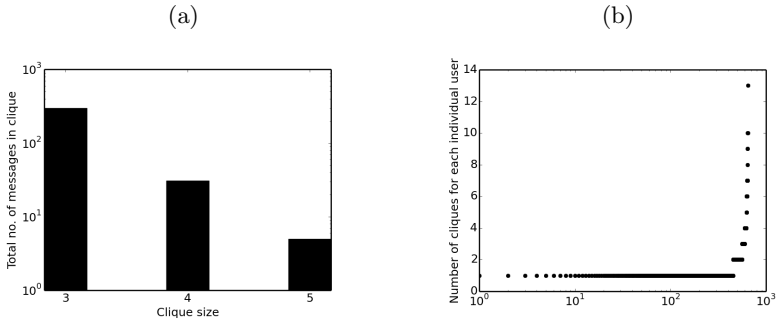

Fig. 5: Panel (a): A size of cliques versus a number of instances (log y-axis). Panel (b):Number of cliques individual users were involved in (log x axis).

6b). We also investigated how balanced multi-user exchanges were, although

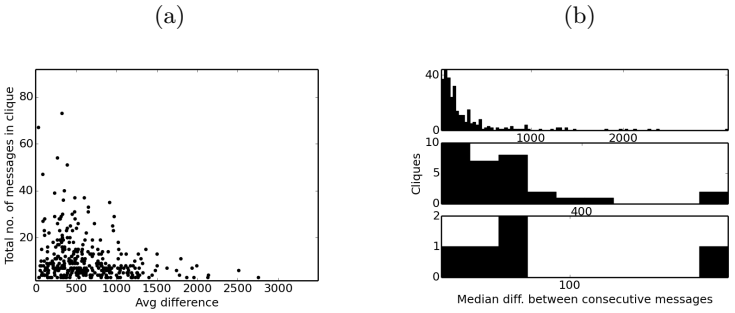

Fig. 6: Panel (a): Average difference in seconds between two consecutive messages in clique versus total number of exchanges. Panel (b): Histogram of medians of differences in seconds between two consecutive messages for cliques of size 3, top, size 4, middle and size 5 bottom.

this situation is more complicated than in the pairwise case.

Firstly, we looked at the difference between the number of tweets received and sent by individual clique members. For each node, we computed the difference of their in-degree and out-degree. We summed up the positive values[8] and to normalise, we divided by the total number of exchanged messages. In this way, we obtained a percentage of 'unreciprocated' messages, where reciprocity is not

[8] Clearly the number of sent and received messages within a group are equal, thus summing the differences between in- and out-degree over individual members in the group is by definition equal to zero.

toward a sender but toward a whole group. We show the histograms for the different sizes of cliques in Fig. 7a. Across all clique sizes and in most of the multi-user conversations around 30% messages were unreciprocated. In a small number of conversations of 3 or 4 users a larger percentage were unreciprocated, i.e. they were dominated by certain members, but also a large number of cliques were very balanced (with unreciprocated messages at $0 - 10\%$), meaning every individual received and sent a similar number of tweets.

Finally, we looked at so-called 'floor-gaining' [4], i.e. how much input each user had over the course of a group exchange[9]. We compared the out-degree of each user within a clique, (remember that each clique is a directed multigraph) with the mean number of edges $r = |n_{\mathrm{E}}|/|n_{\mathrm{V}}|$, where n_{E} is the total number of edges within the clique and n_{V} is the total number of vertices within the clique. In a 'round robin' group conversation, with balanced turn taking, each user would send out r messages, i.e. be responsible for an equal percentage $p = 100r/e$ of the total number n_{E} of exchanged messages. For each clique size, we looked at how many users' representations were greater than or equal to p, i.e. those users who 'dominate' the conversation. On Fig 7b below, we present the histogram for a number of dominant users in the cliques of size 3, 4 and 5. This shows that in

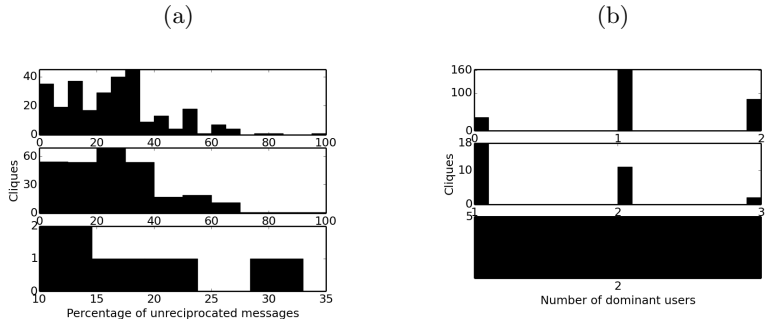

Fig. 7: Panel(a): The percentage of 'unreciprocated' messages for cliques of size 3, top, size 4, middle and size 5 bottom. Panel (b): A number of dominant users in cliques of size 3, top, size 4, middle and size 5 bottom.

most of the cliques of size 3 and 4, one user was responsible for the majority of communication, whilst in cliques of size five, 2 users were dominant. However in about 13% of all cliques of size 3 no users dominated, confirming that Twitter is used for multi-user conversations and not just pairwise conversations.

[9] We argue that the action of tweeting in multiuser exchanges can be regarded as floor-gaining, since tweets with mentions can in principal be read by a wider audience than the group conversing.

4 Conclusions

We looked at conversations in Twitter, based on the underlying structure and timings in approximately 4 million UK tweets with mentions over a period of 28 days. We structured the data as a multigraph to make use of graph algorithms. We proposed a simple method of identifying conversations between pairs of users, based on a time-threshold on the time-to-next tweet, and found evidence that a threshold of 9hrs gives a good indication of distinct conversations. We observed that the conversations detected using this method appeared to be balanced, meaning that each party involved contributed approximately equally to the conversation. This was not the case within more general interactions, in which one agent typically contributed around three times as much as the other.

Although finding cliques in graphs is computationally demanding, because of the sparsity of interactions patterns within the data-set, extracting multi-user exchanges was feasible and relatively fast. We were able to find all cliques within the graph and, using the threshold method, identify conversations for up to a maximum of 5 users. Most of those exchanges were fast-paced. We also found that the number of messages in multi-user exchanges was reciprocal to the average time difference between them. When looking at the balance of multi-user conversations, we found that most exchanges are dominated by just one or two users, with some evidence of well-balanced group exchanges in between 3 users. Regarding the number of received and sent messages by each individual in a group, we found that some were dominated by one or two users, but also some were well balanced.

Further work needs to be done using content information to explore how topics flow through multi-user exchange and if there is any relationship between time-differences between messages and topic. We hope that the insights gained from our analysis could help to develop an understanding of the mechanisms and dynamics of Twitter conversations, with potential scope for generating models of micro-blogging behaviour.

Acknowledgment

This work is partially funded by the RCUK Digital Economy programme via EPSRC grant EP/G065802/1 'The Horizon Hub' and EPSRC MOLTEN EP/I016031/1. We would like to thank Datasift for the provision of the data analysed, and to Colin Singleton and Bruno Gonçalves for very useful feedback and comments.

References

1. L. Backstrom, J. Kleinberg, L. Lee, and C. Danescu-Niculescu-Mizil. Characterizing and curating conversation threads: expansion, focus, volume, re-entry. In *Proceedings of the sixth ACM international conference on Web search and data mining*, WSDM '13, pages 13–22, New York, NY, USA, 2013. ACM.

2. C. Bron and J. Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, Sept. 1973.
3. J. Chen, R. Nairn, and E. Chi. Speak little and well: recommending conversations in online social streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 217–226, New York, NY, USA, 2011. ACM.
4. C. Edelsky. Who's got the floor? *Language in Society*, 10:383–421, 1981.
5. S. Golder and M. W. Macy. Diurnal and seasonal mood tracks work sleep and daylength across diverse cultures. *Science*, 333:1878–1881, 2011.
6. D. Graham. Twitter: value-added conversation is more important than broadcasts, rts. memeburn, August 2012.
7. A. Hagberg, D. Schult, and P. Swart. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference (SciPy2008), Passadena, CA USA*, pages 11–15, August 2008.
8. S. Herring. Computer-mediated discourse. In D. Tannen, D. Schiffrin, and H. Hamilton, editors, *Handbook of Discourse Analysis*. Oxford: Blackwell, 2001.
9. C. Honeycutt and S. C. Herring. Beyond microblogging: Conversation and collaboration via twitter. In *Proceedings of the Forty-Second Hawai'i International Conference on System Sciences, Los Alamitos*, 2009.
10. G. Leboff. Marketing is a conversation? oh yeah ...about what? The Marketing Donut, October 2011.
11. E. Nuutila and E. Soisalon-Soinen. On finding the strongly connected components in a directed graph. *Information Processing Letters*, 49(1):9–14, 1994.
12. D. Quercia, L. Capra, and J. Crowcroft. The social world of twitter: Topics, geography, and emotions. In *The 6th international AAAI Conference on weblogs and social media, Dublin*, 2012.
13. A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *HLT-NAACL 2010*, 2010.
14. D. A. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence: modeling the shape of microblog conversations. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, CSCW '11, pages 355–358, New York, NY, USA, 2011. ACM.
15. J. Sidnell. *Conversation Analysis: An Introduction*. Blackwell, 2010.
16. D. Starkey. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972.
17. R. E. Tarjan. Depth-first search and linear graph algorithms. *SIAM J. Comput.*, 1(2):146–160, 1972.
18. K. Thomas, C. Grier, V. Paxson, and D. Song. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 243–258. ACM, 2011.
19. C. Wang, M. Ye, and B. A. Huberman. From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 244–252, New York, NY, USA, 2012. ACM.
20. S. Williams, M. Terras, and Warwick. What people study when they study twitter: Classifying twitter related academic papers. *Journal of Documentation*, 69, 2013.
21. S. Wu, J. Hofman, W. Mason, and D. Watts. Who says what to whom on twitter. In *WWW, 2011*, 2011.
22. S. Wu, C. Tan, J. Kleinberg, and M. Macy. Does bad news go away faster? In *5th International AAAI Conference on Weblogs and Social Media, 2011*, 2011.