

On the Study of Social Interactions in Twitter

Sofus A. Macskassy

Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
sofmac@isi.edu

Abstract

Twitter and other social media platforms are increasingly used as the primary way in which people speak with each other. As opposed to other platforms, Twitter is interesting in that many of these dialogues are public and so we can get a view into the dynamics of dialogues and how they differ from other other tweet behaviors. We here analyze tweets gathered from 2400 twitter streams over a one month period. We study social interactions in three important dimensions: what are the salient user behaviors in terms of how often they have social interactions and how these interactions are spread among different people; what are the characteristics of the dialogues, or sets of tweets, that we can extract from these interactions, and what are the characteristics of the social network which emerges from considering these interactions? We find that roughly half of the users spend a fair amount of time interacting whereas 40% of users do not seem to have active interactions. We also find that the vast majority of active dialogues only involve two people despite the public nature of these tweets. We finally find that while the emerging social network does contain a giant component, the component clearly is a set of well-defined tight clusters which are loosely connected.

The use of “micro-blogging” services, such as Twitter, has exploded exponentially in recent years. For example, currently, millions of Twitter users post millions of 140-character messages, called “Tweets,” about topics ranging from daily activities, to opinions, to links to funny pictures. Beyond the large collection of user generated text, Twitter also has a social network aspect, allowing users to publicly message one another directly, and set up a social network of people who follow one another’s Tweets. This rich relational and textual setting has spurred research in a number of areas beyond traditional network analysis (e.g., (Kwak et al. 2010; Krishnamurthy, Gill, and Arlitt 2008)). For instance, Twitter has been analyzed to discover breaking news (Sankaranarayanan et al. 2009), as a forum for analyzing media events (Shamma, Kennedy, and Churchill 2009), as a vehicle for information diffusion (Leskovec et al. 2007; Lerman and Ghosh 2010; Lerman and Hogg 2010), as a mechanism for language learning (Borau et al. 2009), and even for detecting

natural disasters in real-time (Sakaki, Okazaki, and Matsuo 2010).

Much recent work in microblogs as described above tend to treat the social media streams and underlying social networks as large global phenomena where global processes, metrics and statistics rule the day. In other words, the streams, people and links in these social media are all treated as a large homogeneous mass. While such a high-level view of the world is of tremendous use in order to understand large global behaviors, it unfortunately is not appropriate for fine-grained analysis of local behaviors. For example, community detection fails to find meaningful clusters on these large networks (e.g., (Leskovec et al. 2008)), information diffusion and other metrics match on macro-level but fails to fit observed data at the micro-level. Some recent work does look at more local behaviors such as behaviors of retweeting (e.g., (Nagarajan, Purohit, and Sheth 2010; Macskassy 2011)).

This paper also focuses on local behaviors, particularly that of public chatter within Twitter. In particular, we seek to better understand online chatting behavior in a forum such as Twitter, both in terms of time/attention spent on chatter, whether chatter looks fundamentally different from other types of tweets, and what kind of social networks are formed through chats. In particular, do we see large connected and dense components forming such as what we have seen in the blogosphere (e.g., (Kumar, Novak, and Tomkins 2006)), or do people tend to stay chat with a smaller and tighter group of people?

The key to our contribution lies in our analysis of thousands of automatically extracted dialogues from over 650,000 tweets collected by monitoring over 2,400 Twitter users for a period of one month. We extracted the dialogues, and analyzed who talks to whom, how often the users chat versus other activities (e.g., retweeting or tweeting general information). Our analysis shows some interesting behaviors and verifies some common sense beliefs about how users interact on Twitter. We find that roughly half of the users spend quite a bit of time chatting, that chat-type tweets are indeed different from other tweets (generally shorter) and that most dialogues tend to occur between pairs of people. We also find that although large components do form, they tend not to be dense as seen elsewhere (e.g., (Leskovec et al. 2008)), but rather there are large clusters which are tied through a

few people. We also find that chatter in general is very fair in that users receive as much as they give and that they often have very equitable participation (e.g., it is not the case that one user generally is the dominant speaker with most others being silent).

The rest of the paper is outlined as follows: we start by discussing related work. We then describe the process we use to extract dialogues from Twitter streams, followed by a case study on our Twitter data. We finish with a discussion of our findings.

Related Work

Of most relevance to this paper are recent explorations of users, such as how they split their attention across friends (Lars Backstrom and Rosenn 2011), modeling discourse (Strzalkowski et al. 2010) and chat (Dong2006 2006), the roles users play in online forums (Welser et al. 2007), and the underlying processes by which they decide to retweet information (Nagarajan, Purohit, and Sheth 2010; Macskassy 2011). All of these work center on the personal user behavior or try to disentangle dialogues from streams. In this paper we focus a little more on what the general discourse behaviors look like in a public broadcast forum such as Twitter. In this case we have explicit dialogue structure through the “@user” construct, so we do not need to disentangle threads. Also here, we are more interested in meta-behaviors such as how much time is spent chatting as well as how the emerging dialogue social networks are structured.

There is also a rich literature on the general exploratory mining of the blogosphere (e.g., (Joshi et al. 2007; Leskovec et al. 2007; Agarwal and Liu 2008; Hearst and Dumais 2009)). Much has been said about the the underlying topology (Shi, Tseng, and Adamic 2007), demographics (Kumar et al. 2003), structure (Kumar, Novak, and Tomkins 2006) and evolution (Kumar et al. 2003; Backstrom et al. 2006; Kumar, Novak, and Tomkins 2006; Leskovec, Kleinberg, and Faloutsos 2007; Götz et al. 2009) of the blogosphere.

Analyzing the temporal nature of the social network is clearly important. In addition to the broad work on evolution above, one can also focus on the individual level to understand how individuals form and break bonds to communities (e.g., (Sharara et al. 2010)). Improving capabilities in that respect would greatly help understanding the dynamic nature of how people interact with each other.

More recently, researchers have turned towards better understanding of how information flows through the blogosphere. The methodology used is generally one of understanding information cascades (Leskovec et al. 2007; Papagelis, Bansal, and Koudas 2009; Ghosh and Lerman 2011). In other words, what are the specific patterns of diffusion for specific pieces of information. These are generally relatively small patterns, but still informative in order to understand at the micro-level how information might be passed along.

Community detection algorithms have received significant attention in recent years (see, e.g., (Clauset, Newman, and Moore 2004; Newman 2005; Leskovec et al. 2008; Porter, Onnela, and Mucha 2009)). The most common approaches take a graph (such as a social network) and split it

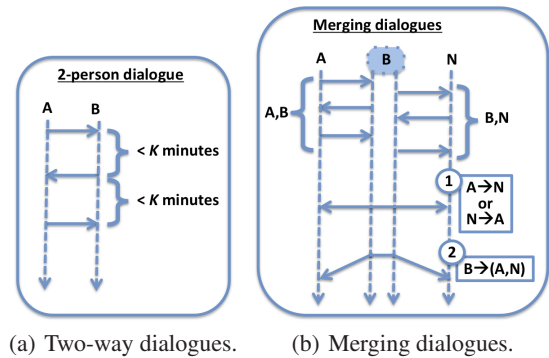


Figure 1: Process for extracting dialogues from Twitter.

into k disjoint clusters, where each cluster supposedly represents a “community” in that graph.

Extracting Dialogues from Twitter Streams

Key to our analysis of how people interact or chat in Twitter is how we extract the dialogues from the Twitter streams. In our case we are interested exclusively on explicit dialogues where users specifically mention each other using the “user” construct. These tweets are specifically addressed to these other users. In order for there to be a dialogue, however, the other user must “reply” back by sending a tweet mentioning the original user. Clearly, for there to be a live dialogue these have to happen within a short period of time (say, within K minutes).¹ We further restrict our definition of a dialogue to only include dialogues where there are at least three tweets and at least two active users (for example, if one user sends three tweets to another users within the time window, but never receives a reply, then that does not count as a dialogue). We show the process of a minimal prototypical two-way dialogue in Figure 1(a) and provide a concrete example from our data in Table 1.

Next, it may well be that users talk to more than one other user at a given time. We will normally consider these as two separate concurrent dialogues, except in the case where all the users are clearly speaking with each other. This can happen either because multiple users are referred to in the same tweet (i.e., multiple uses of “user”, or because the separate users start speaking with each other (closing the triangle). The overall merging process is visualized in Figure 1(b) and a concrete example from our data is shown in Table 2.

Case Study

We now turn to our case study. We focus here on a Twitter data set which we have processed using the approach above to generate dialogues. We here explore the emergent behaviors of dialogues in the Twittersphere. Specifically, we look at aggregate user behavior, their dialogues as well as the social networks formed by user interactions. We therefore will first need to process the data into dialogues and networks.

¹We discuss selecting K below in the Section describing the Case Study.

Time	User	Tweet
23:56	User1	@User2 why don't you get a car my friends
00:00	User2	@User1 cause my cars transmission blew before i left remember..i may get a new one when i come back
00:01	User1	@User2 ohh and when you come back we must go to chipotle together when is that?
00:03	User2	@User1 DEFINITELY im going there and in an out every dayyyy
00:09	User1	@User2 do you not have a in n out to??
00:16	User2	@User1 no we do but its hella far :(i come back in december
00:19	User1	@User2 my birthday!! I'll drive??
00:23	User2	@User1 im sooo down..my parents wanna get me a convertible bug lol

Table 1: Example two-way dialogue.

ID	Time	User	Tweet
(A)	18:55	User1	@User7 MQM is THE MAFIA! The organized crime in Karachi! Now, please! @User2
(A)	18:57	User1	@User2 bro, please stop misstating me. I love Karachi and the people of Karachi. But ...
(A)	19:00	User2	@User1 Mafia of MQM makes 70% of Karachi, then wht do U luv hre? The remaing 30%? ...
(B)	19:01	User3	@User4 @User5 @User2 v hope 4 a political d judical systm in our country! whoevr fulfil ...
(A)	19:04	User1	@User2 70% of Karachi is MQM? Really?? Is that how you learn your other 'facts,' too? :) ...
(C)	19:05	User2	@User6 @User1 @User5 @User3 MQM waloon ki Qabar Karachi mai hi hoti hai. Kon apnay ...
(C)	19:06	User2	@User1 ok then app battaa do... Laikin baat tou puri karo ...
(C)	19:09	User4	@User3 @User5 @User2 :) No Maseeha other than Imam Mehdi&ESA(AS) will come, ...
(C)	19:10	User2	@User5 @User6 @User1 @User3 wrong, unn ki maiyat gaoo jaati hai, as per tradition. There...
(C)	19:11	User2	@User4 @User3 @User5 no doubt about that, but until then we make way.
(C)	19:13	User1	@User2 :) Your 'facts' tell me this discussion would go nowhere. Besides, ...

Table 2: Example of merging two dialogues. User2 was active in dialogue (A) and implicit in dialogue (B) as he was mentioned. The user tied the dialogues together by “mentioning” users from both of these dialogues.

From here, we then analyze each of these three artifacts. We next describe our data in some detail, followed by a discussion of our data processing, and then we present the results of our analysis.

Data

We have been continuously collecting Tweets from a set of about 2,400 Twitterers in the Middle East to explore a geographically constrained set of individuals. We identified these individuals using a snowball sampling method where We started from a seed set of ~125 Twitterers who self-reported (in their profile) to reside in the Middle East. From there, we expanded the set of users to monitor whenever we saw a retweet or a mention (the *user* construct), adding only users who self-reportedly were in the same region. After two iterations we had had reached ~2,400 Twitterer users, which is the set of users we study in this paper. The snowball sampling has yielded a constrained set of users who make up on large connected component. It is not unreasonable to assume that we have a slice of Tweets that many of them are aware of, or at the very least is representative of a geographically focused set of Tweets they are likely to see. Clearly this is a geographically biased sample, but it is also powerful because it is thusly constrained and is therefore quite useful as a deeper study in a geographic region.

We have been monitoring these Twitterers and any mention of them since early September 2010, and have tagged the first month of data (9/20/2010 through 10/20/2010),

which makes up our data set. The full tweet data set for this period consists of 652,600 Tweets, 498,056 of which were tweeted by our 2,400 users (the remaining being mentions of one or more of our users). We include mentions of the users here because we are studying the dialogues which they participate in, which could include users not in our data set. In fact, the full set contains over 30,000 twitter users.

Experimental Methodology

The case study we present in this paper is focused on three dimensions of dialogue behaviors in Twitter:

1. **Users:** How do users behave? How many tweets are in a dialogue (vs. tweeting or retweeting), how many people do they speak to, how many dialogues do they have, etc.
2. **Dialogue** itself: How long are dialogues, how many people participate, and can we say something about the dialogue graph.
3. **Social Network:** If we consider who speaks to whom over time in these dialogues, what does the resulting social network look like? Is it fragmented, does it contain clear clusters, ... ?

User Behaviors First, we are interested in how users behave. Specifically, how many of their tweets are spent on dialogues vs. other tweets? We categorize tweeting into four types:

1. **Dialogue:** This tweet is part of an active dialogue.

2. **Mention:** This tweet mentions another user but is not in a dialogue.
3. **Retweet:** This is a retweet.
4. **Tweet:** Remainder of tweets.

We first want to understand the general activity behavior of users in terms of the kind of tweets they post. Furthermore, we are interested in understanding how many tweets with a “mention” starts a new dialogue (conversion rate), how well-connected a user is (degree) and whether the connectivity is one-sided or both (e.g., is it the case that the user tends to mention others, being mentioned by others or is it fairly even). In terms of connectivity and activity, do users tend to have a few strong connections or do they spread their social interactions across more users? To this end, we compute the **chat-entropy** of a user. If we consider the activity of a user, we can ask how much “time” the user spends with another user. For example, if the user talks to 10 people in 5 dialogues, how many dialogues included “user1”? or how many dialogue tweets mentioned “users1”? We compute the entropy along these two dimensions to get an understanding of users in their social interactions. We compute the dialog-based chat-entropy as follows:

$$\text{Ent}_D(u) = - \sum_{n \in \mathcal{N}_u} \left(\frac{d_{u,n} * \log(d_{u,n})}{+(1 - d_{u,n}) * \log(1 - d_{u,n})} \right),$$

where \mathcal{N} is the set of users that u has tweeted to in one or more dialogues and $d_{u,n}$ is the ratio of u ’s dialogues which included user n . A *high entropy* in this metric suggests that u speaks with each friend roughly half of the time whereas a low entropy suggests that the user mostly have dialogues with the same people over and over again.

Similarly, we compute the tweet-based chat-entropy as follows:

$$\text{Ent}_T(u) = - \sum_{n \in \mathcal{N}_u} \left(\frac{r_{u,n} * \log(r_{u,n})}{+(1 - r_{u,n}) * \log(1 - r_{u,n})} \right),$$

where \mathcal{N} is defined as above and $r_{u,n}$ is the ratio of u ’s tweets in dialogues which mentioned user n . In this metric, a *low entropy* means that a user tends to speak to only a few of her “neighbors” whereas a *high entropy* means that the user more evenly distributes her attention to all users she speaks with.

Finally, we note that in order to compare entropies across dialogues and sets of users of different sizes, we normalize the entropy score to force it to lie between 0 and 1. This normalizing constant is: $(-|n| * \frac{1}{2} * \log(\frac{1}{2}))$, where $|n|$ is the size of the set.

Dialogues We described in a prior Section how we extract dialogues from the Twitter stream. The only parameter we have is the amount of time between chat-tweets we allow to consider two people actively engaged in a dialogue. We explored different settings, from one minute to nine minutes to understand how this would change the underlying structure. While this is ultimately subjective (i.e., if users stop for 8 minutes and then begins talking again on the same topic, is that a single dialogue or two dialogues?). We discuss choosing a time window in the case study below.

Tweet Category	Number Tweets	Overall Ratio
Dialogue	66,812	0.13
Retweet	93,319	0.19
Mention	154,177	0.31
Tweet	183,748	0.37
Total	498,056	1.00
Conversion	20,155	0.12

Table 3: Overall break-up of Tweets over our data set. The last row shows the total number of dialogues and the ratio of tweets with a “mention” (outside active dialogues) which resulted in a new dialogue being started.

Once we have extracted a dialogue, we calculate the following metrics:

1. **Size:** What are the number of *active* participants in the dialogue? If someone is mentioned but never replies, then that user is not considered part of that dialogue).
 2. **Chat-Entropy:** If we consider all the active users and how much they participate (the number of tweets), how spread out is the dialogue? For example, is one user dominating or is the overall chatter evenly spread out? We compute the entropy of chats in the following way:
- $$\text{Ent}(D) = - \sum_{u \in D} \left(\frac{r_{u,D} * \log(r_{u,D})}{+(1 - r_{u,D}) * \log(1 - r_{u,D})} \right),$$
- where $r_{u,D}$ is the ratio of tweets in dialogue D which originated from user u . High entropy means that all users participate evenly, whereas a low entropy means a few users dominate. We normalize this entropy score as well for comparative purposes in the analysis.
3. **Density:** As the number of users go up, does the density of dialogue graph go down?
 4. **Reciprocity:** As the number of users go up, how much reciprocity in mentions to we observe?

Social Network Once we have extracted dialogues and active interaction between users, then we form a social network of all these interactions. The questions we seek to answer on these networks are whether the emerging network is connected or fragmented. Do we observe a giant component forming? If so, is this a dense component or do we see clear communities within that component? If so, are there clear “ambassadors” who connect multiple communities? Is there a clear recursive structure (e.g., communities within communities?)

Analysis

We present in this section our analysis of Twitter dialogues across the three dimensions we described above. As mentioned above, one key aspect to the overall analysis is setting k , the time-window parameter for extracting dialogues. We use $k = 5$ for our analysis. We note that we did explore other values of k identical qualitative results and we omit discussing those here due to space constraints.

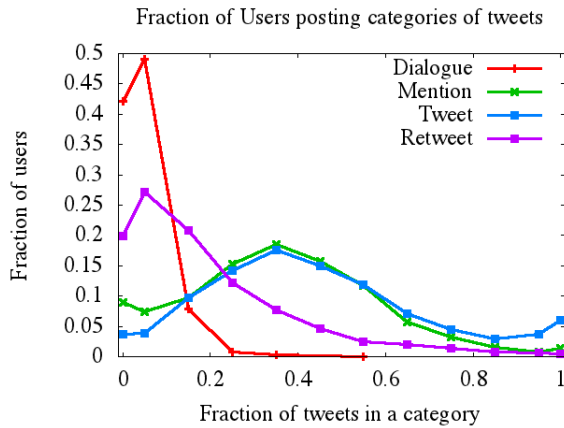


Figure 2: Activity profiles of users: what fraction of users spend 0% through 100% of their time posting each type of tweet?

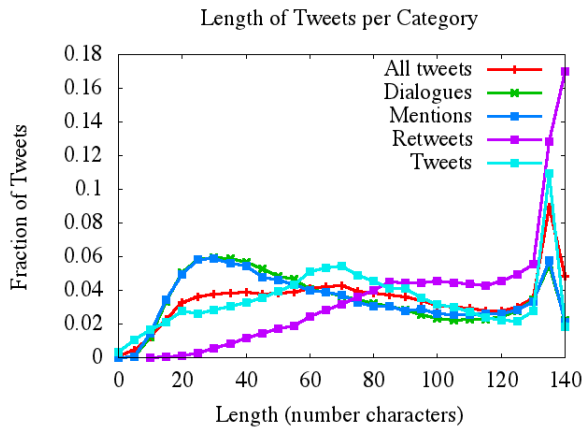


Figure 3: Length of tweets across categories.

User behaviors We first explore user behaviors and the first question we have is how often users perform various activities. Table 3 shows the aggregate statistics on the number of tweets in different categories and how often a tweet lead to a dialogue.

As we can see, dialogues make up about 13% of user activity. However, it turns out that 42% of users did not participate in any dialogues, so this number may be skewed. We therefore turn looking at the aggregate distribution of tweets over all users. Figure 2 shows what fraction of users spend $x\%$ of their activity on a particular tweet category. For example, the figure shows that nearly 20% of users spend 35% of their online activities on “Tweeting” and “Mention” tweets, that 50% of users spend 5% of their time in dialogues a fair amount of users (20%) do not retweet at all, while 10% of users only tweet.

In addition to the amount of tweets per category, we also explored whether the different types of tweets had other different characteristics. Of particular interest was the length

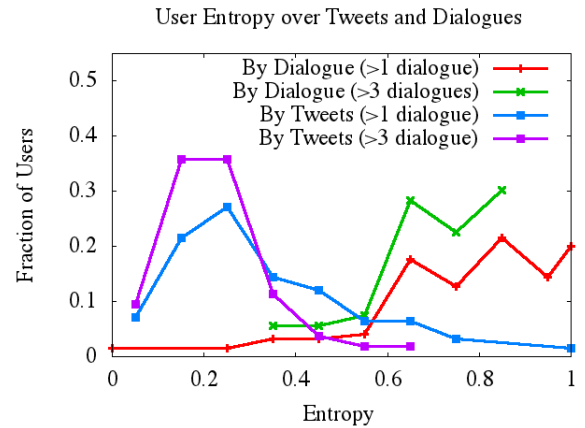


Figure 4: How do users break up attention?

of the tweets. Figure 3 shows the distribution of lengths of tweets per category. We see significant differences, where tweets with mentions—direct interaction tweets—clearly are much shorter than any of the other tweets. We also see that retweets generally are much longer, although this is probably biased due to the added characters to indicate it is a retweet.

The key question we had for users was where they put their attention and whether they tended to talk to a few people or more people. While the vast majority of people only had two-way dialogues, we wanted to know whether they always spoke with the same people. We also wanted to know whether, if they spoke with multiple people, they would spread their conversation evenly or focus on a few select close friends. To this end, we computed the chat-entropy for each user. Figure 4 shows the distribution of entropies over all users, focusing on users who had at least 2 dialogues and also on users who had at least 4 dialogues. The figure shows both types of chat-entropies for comparative purposes and the differences are striking. First, we note that the chat-entropy by dialogue is very high, suggesting that users tend to mix and match who they speak to rather often, which holds for people having at least 4 dialogues. However, the chat-entropy conditioned on tweets show that while the user may speak with different people across different dialogues, most of the volume is directed at the same people repeatedly. This is interesting because it does suggest that even if a user participates in different dialogues with multiple people, they still tend to direct their tweets to their closer friends.

Our final analysis of user behavior looks closer at the interactions of users over time—for pairs of users, does one tend to speak more than the other? We computed for each dialogue tweet, the number of times users tweeted to each other over the lifetime of this study. This resulted in a weighted graph, where a directed edge from a to b had the weight equivalent to the number of times a tweeted to b . From here we compute the weighted in-degree and out-degree for each pair of users who had at least one interac-

Size	Number	Ratio	Avg. Num. Tweets
2	18,619	92.37%	4.9
3	1,232	6.11%	8.5
4	181	0.90%	12.7
5	83	0.41%	19.4
6	27	0.13%	36.5
> 6	13	0.07%	> 60

Table 4: High-level statistics of extracted dialogues.

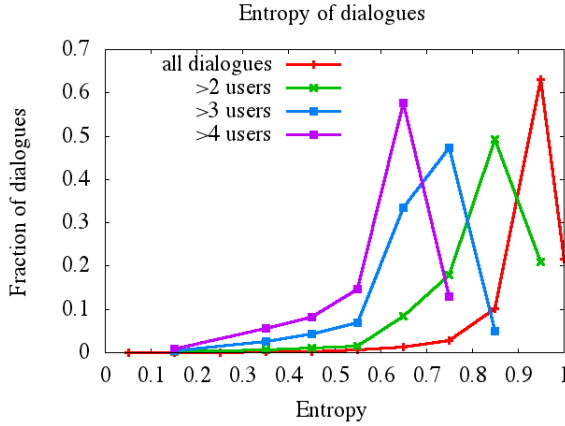


Figure 5: Chat-entropy of dialogues.

tion. These are highly correlated (Pearson’s Correlation of 0.9978), showing that dialogues were very well balanced.

Dialogues Our next area of study is the dialogues themselves. We wanted to know how many dialogues we extracted and the characteristics of those dialogues. Table 4 shows that the vast majority of dialogues (92%) extracted were between two people. We also see that as the number of users increased, so do the average number of tweets. However, the dialogues were still relatively short—less than 5 tweets total for the most part.

We first wanted to know how balanced dialogues are: is the conversation generally even across all users, or is it the case that a few people tend to dominate. We computed the chat-entropy for a dialogue to get insight into this question. Figure 5 shows that generally the entropy is quite high (well-balanced discussions), but that this goes down as we down-select to larger and larger groups. This suggests that a few people tend to lead discussions more in larger groups.

Next, we looked at the density. The group sizes for our dialogues are often relatively small (6 and below), so it would not be unreasonable to hypothesize that dialogues consisting of such a small number of users would be fairly dense (e.g., everybody directs a tweet to each other at least once during the dialogue). Figure 6 shows how the density of the dialogue graphs change as we condition on the number of active participants. The figure clearly shows that small groups have a high density but as we only consider larger groups, the density drops rapidly, even when there are only 3 people

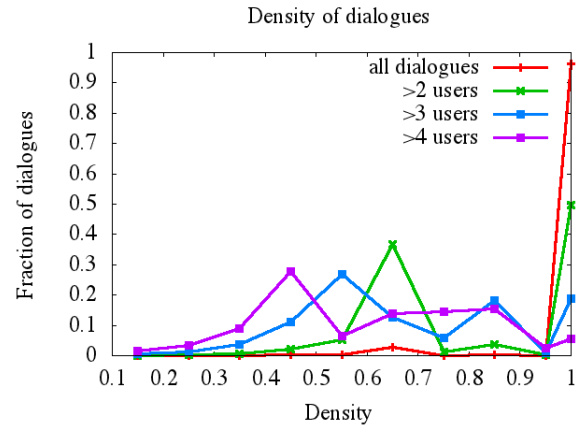


Figure 6: Density distributions of dialogues.

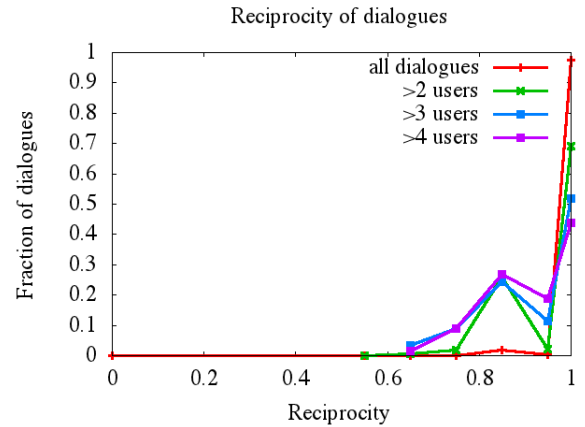


Figure 7: Reciprocity distributions of dialogues.

involved.

Given that the dialogue graphs were not fully connected, we wanted to know whether directed tweets were reciprocated. Normal etiquette would suggest that people ought to reciprocate a directed tweet during a conversation. The way we define reciprocity in this study is: if there is one directed tweet from a to b in this dialogue, do we also see at least one directed tweet from b to a . We do not consider the number of directed tweets. Figure 7 shows how reciprocity changes as the group sizes increase. Interestingly enough, we see that although we have very high reciprocity rates, they do go down significantly as the group size goes up. For example, only half of the dialogues having more than 3 active participants had 100% reciprocity.

Network Structure Our last analysis of our dialogues is on the emerging social network in which these dialogues take place. As we extract the social network from these personal interactions, we want to know whether we see similar behavior as has been reported elsewhere in terms of giant and dense components emerging.

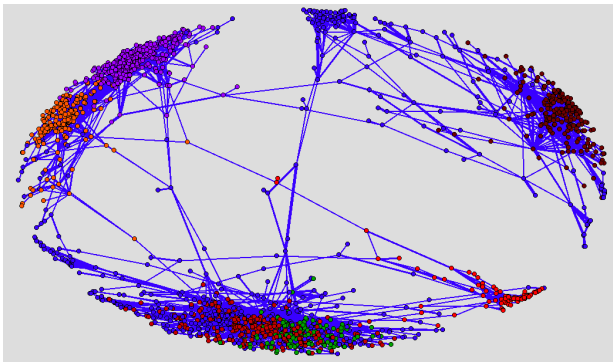


Figure 8: Graph structure of largest component, colored by major clusters found using modularity clustering. We see clear structures in the graph, where the clusters are loosely connected.

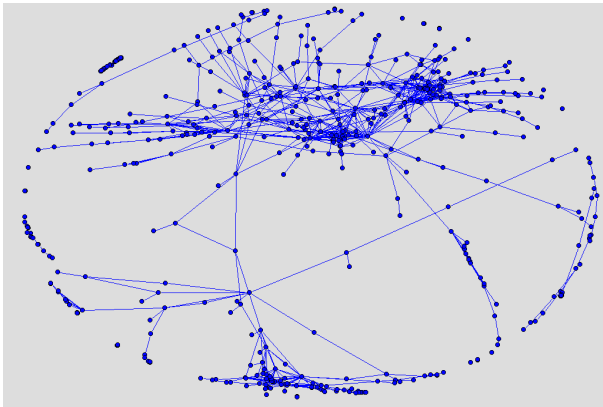


Figure 9: Graph structure after removing the largest clusters. As with the full component, was still see strong structures, and even a few disconnected components.

The resulting social network consisted of 5952 users who were actively participating in dialogues, with a total of 19838 edges. The resulting network did in fact follow similar high-level characteristics found elsewhere, where we found one giant component of 4609 users, following by component of size 231, 110 and then 46 and below. There were a total of 179 components.

We first wanted to know whether this giant component was dense or whether it was a set of loosely connected tighter communities. We first ran modularity clustering on the giant component, which resulted in 55 clusters, where the first five clusters ranged from sizes of 300 to 978. We show the component in Figure 8, where we provide different colors to the six largest clusters, and giving a single color to the remaining clusters. We immediately see that the component is actually quite loosely connected with respect to those clusters. We removed the nodes from these large clusters and laid out the remainder of the nodes in Figure 9. We still see the same general structure, although we now have a couple of disconnected components as well. Clearly we have the case where there are larger well-defined communities in

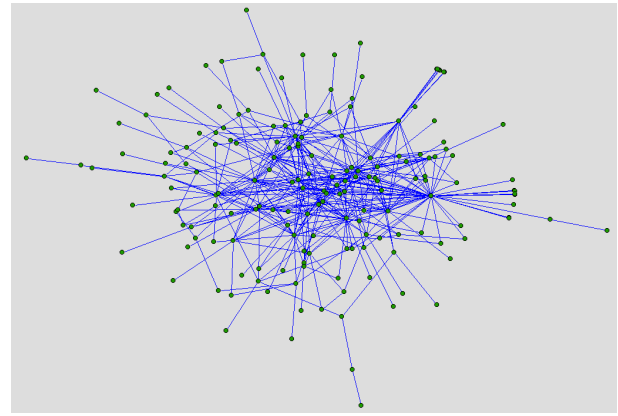


Figure 10: Graph structure for one of the large clusters found in the giant component. The other large clusters follow similar structural characteristics.

the giant component. We also note that there does not seem to be single people connecting these communities—rather we see that they are connected through a string of intermediates.

Finally, we wanted to explore the clusters themselves to see if perhaps we would find a recursive structure with clear smaller communities within these clusters. Figure 10 shows one of the clusters where we see no strong indicator of tightly connected smaller communities. All the other large clusters share a same characteristic.

Discussion

This paper focused on exploring in some detail the behavior and dynamics of dialogues in Twitter. In particular, we wanted to user interaction behaviors, the characteristics of the dialogues people where having and the structure of the emerging social network generated by these interactions.

We found that most people either do not have dialogues or spend about 5 – 10% of their Twitter activity in direct interaction with other users. We also found users are very equitable in their interactions, giving and receiving in equal amounts. Interestingly, we found that users were in dialogues with different subsets of people but still tended to directly interact with only a few.

We found that the vast majority (over 92%) of dialogues were between two people, about 6% of dialogues were between three people with marginal fractions for larger groups. Despite this, we saw a very strong trend for dialogues involving larger number of people tended to not be well-connected although reciprocity was always very high.

Finally, we observed similar emerging social network behavior reported elsewhere, where a giant component was forming out of the social interactions. However, we further noted that this giant component clearly was a set of well-defined smaller communities which were loosely connected.

References

- Agarwal, N., and Liu, H. 2008. Blogosphere: Research issues, tools, and applications. *SIGKDD Explorations* 10(1):18–31.

- Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*.
- Borau, K.; Ullrich, C.; Feng, J.; and Shen, R. 2009. Microblogging for language learning: Using twitter to train communicative and cultural competence. In *Proceedings of the International Conference on Advances in Web Based Learning*.
- Clauset, A.; Newman, M. E. J.; and Moore, C. 2004. Finding community structure in very large networks. *Physical Review E* 70. 066111.
2006. Structural analysis of chat messages for topic detection. *Online Information Review* 30(5):496–516.
- Ghosh, R., and Lerman, K. 2011. A framework for quantitative analysis of cascades on networks. In *Proceedings of Web Search and Data Mining Conference (WSDM)*.
- Götz, M.; Leskovec, J.; Mcglohon, M.; and Faloutsos, C. 2009. Modeling blog dynamics. In *AAAI Conference on Weblogs and Social Media (ICWSM)*.
- Hearst, M., and Dumais, S. 2009. Blogging together: An examination of group blogs. In *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media*.
- Joshi, A.; Finin, T.; Java, A.; Kale, A.; and Kolari, P. 2007. Web 2.0 Mining: Analyzing Social Media. In *Proceedings of the NSF Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation*.
- Krishnamurthy, B.; Gill, P.; and Arlitt, M. 2008. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*.
- Kumar, R.; Novak, J.; Raghavan, P.; and Tomkins, A. 2003. On the bursty evolution of blogspace. In *Proceedings of the 12th international conference on World Wide Web*.
- Kumar, R.; Novak, J.; and Tomkins, A. 2006. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*.
- Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the International Conference on World wide web*.
- Lars Backstrom, Eytan Bakshy, J. K. T. M. L., and Rosenn, I. 2011. Center of attention: How facebook users allocate attention across friends. In *International Conference on Weblogs and Social Media (ICWSM)*.
- Lerman, K., and Ghosh, R. 2010. Information contagion: an empirical study of spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*.
- Lerman, K., and Hogg, T. 2010. Using a model of social dynamics to predict popularity of news. In *Proceedings of 19th International World Wide Web Conference (WWW)*.
- Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007. Cascading behavior in large blog graphs. In *SIAM International Conference on Data Mining (SDM)*.
- Leskovec, J.; Lang, K.; Dasgupta, A.; and Mahoney, M. 2008. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. arXiv:0810.1355v1.
- Leskovec, J.; Kleinberg, J.; and Faloutsos, C. 2007. Graph evolution: Densityification and shrinking diameters. *ACM Trans. Knowl. Discov. Data* 1.
- Macskassy, S. A. 2011. Why do people retweet? anti-homophily wins the day! In *International Conference on Weblogs and Social Media (ICWSM)*.
- Nagarajan, M.; Purohit, H.; and Sheth, A. 2010. A qualitative examination of topical tweet and retweet practices. In *International Conference on Weblogs and Social Media (ICWSM)*.
- Newman, M. 2005. Modularity and community structure in networks. In *Proceedings of the National Academy of Sciences*, 8577–8582.
- Papagelis, M.; Bansal, N.; and Koudas, N. 2009. Information cascades in the blogosphere: A look behind the curtain. In *AAAI Conference on Weblogs and Social Media (ICWSM)*.
- Porter, M. A.; Onnela, J.-P.; and Mucha, P. J. 2009. Communities in networks. *Notices of the AMS* 56(9):1082–1097, 1164–1166.
- Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the International Conference on World Wide Web*.
- Sankaranarayanan, J.; Samet, H.; Teitler, B. E.; Lieberman, M. D.; and Sperling, J. 2009. Twitterstand: news in tweets. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.
- Shamma, D. A.; Kennedy, L.; and Churchill, E. F. 2009. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*.
- Sharara, H.; Singh, L.; Getoor, L.; and Mann, J. 2010. Understanding actor loyalty to event-based groups in affiliation networks. *Journal of Advances in Social Networks Analysis and Mining*.
- Shi, X.; Tseng, B.; and Adamic, L. 2007. Looking at the blogosphere topology through different lenses. In *AAAI Conference on Weblogs and Social Media (ICWSM)*.
- Strzalkowski, T.; Broadwell, G. A.; Stromer-Galley, J.; Shaikh, S.; Taylor, S.; and Webb, N. 2010. Modeling socio-cultural phenomena in discourse. In *Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10*.
- Welser, H. T.; Gleave, E.; Fisher, D.; and Smith, M. 2007. Visualizing the signatures of social roles in online discussion groups. *The Journal of Social Structure* 8(2).