Network Effects on Tweeting

Jake T. Lussier and Nitesh V. Chawla

Interdisciplinary Center for Network Science and Applications (iCeNSA),
University of Notre Dame, Notre Dame, IN 46556, USA
{jlussier,nchawla}@nd.edu
http://www.nd.edu/~(jlussier,nchawla)/

Abstract. Online social networks (OSNs) have created new and exciting ways to connect and share information. Perhaps no site has had a more profound effect on information exchange than Twitter.com. In this paper, we study large-scale graph properties and lesser-studied local graph structures of the explicit social network and the implicit retweet network in order to better understand the relationship between socialization and tweeting behaviors. In particular, we first explore the interplay between the social network and user tweet topics and offer evidence that suggests that users who are close in the social graph tend to tweet about similar topics. We then analyze the implicit retweet network and find highly unreciprocal links and unbalanced triads. We also explain the effects of these structural patterns on information diffusion by analyzing and visualizing how URLs tend to be tweeted and retweeted. Finally, given our analyses of the social network and the retweet network, we provide some insights into the relationships between these two networks.

Keywords: Data mining, social networks, information diffusion.

1 Introduction

A microblogging service launched in 2006, Twitter.com has since grown into one of the most popular OSNs on the web and one of the most prominent technological influences on modern society. Twitter is not only an environment for social tie formation, but also for information exchange and dissemination. Whether it be trivial status updates from casual users, news postings from media and blogging services, or logistical instructions from citizens organizing political movements, Twitter provides a truly unique environment for diverse forms of information sharing. As such, many recent studies have focused on this content. For example, in [8], Romero et al analyze the diffusion of different hashtags, find variation due to "stickiness" and "persistence," and observe that political tweets are by far the most persistent. Castillo et al employ a supervised classification scheme to assign tweet and credibility labels to user tweets [3]. Other papers have taken various approaches to tweet sentiment analysis, including the study performed by Dodds et al that reports a steady global happiness level with weekly and daily patterns [4].

T. Elomaa, J. Hollmén, and H. Mannila (Eds.): DS 2011, LNCS 6926, pp. 209–220, 2011.

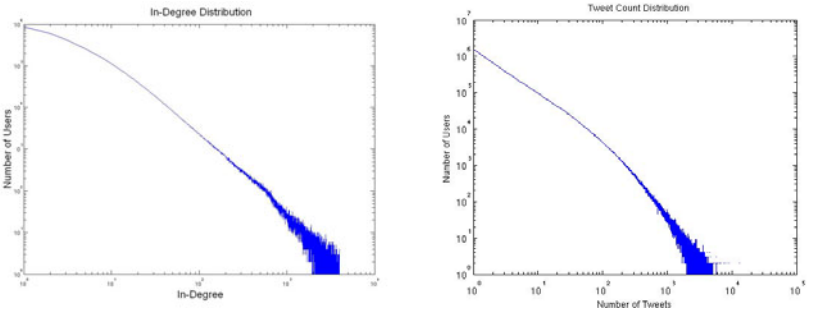
Beyond studying tweets content, much work has focused on understanding how the importance of information sharing affects social network properties and phenomena. As Kwak et al report in [5], the Twitter network diverges from other social networks in that "its distribution of followers is not power-law, the degree of separation is shorter than expected, and most links are not reciprocated." More recently, Wu et al investigated "Who Says What to Whom on Twitter" [10] and found that 50% of URLs consumed by Twitter users are tweeted by just 20 thousand "elites." Moreover, by differentiating users, organizations, media, and bloggers, this paper studied how these categories differ from one another and how information flows amongst them. In doing so, Wu et al find evidence of homophily within groups and offer support for the "Two-Step Flow" theory from communications theory.

In addition to studying the macro properties of the Twitter social / information network, other work has focused on the behaviors of individual users, focusing especially on measuring user influence. In [1], Bakshy et al report that the largest information cascades are generated by users who were previously influential and who post "interesting" content. Welch et al show in [9] that PageRank scores based on the implicit retweet network are more reliable than those based on the explicit social network.

In this paper, we leverage macro and local analyses of the explicit and implicit Twitter networks in order to better understand the relationship between social ties and user tweets. More specifically, we begin in Section 2 by explaining our data sources and some simple network statistics. We then study the interplay between the explicit social network and tweet topics in Section 3 and provide evidence suggesting that proximity in the social graph generally corresponds to higher tweet similarity. We next investigate properties and behaviors of the implicit retweet network, reporting interesting properties in Section 4 and then describing and visualizing local graph structures that arise around the diffusion of URLs in Section 5. Finally, we directly study the relationship between the social network and the retweet network in Section 6. We conclude in Section 7 by summarizing our findings and discussing future research.

2 Data

In this paper, we use data from two sources. First, we obtained social network data from researchers at KAIST (data available at `an.kaist.ac.kr/traces/WWW2010.html`). This social network, which was initially analyzed by Kwak et al in [5], lists 1.47 billion social ties and 41.7 million users. Second, we obtained tweet data from researchers at Stanford University. This dataset consists of 50% of all tweets sent during a seven month period and includes tweet time, author, and text. We then filtered these two datasets so that each only contains social ties and tweets from users who have appeared in both datasets. The resulting social network degree distribution can be seen in Figure 1a and the node-tweet count distribution can be found in Figure 1b.



(a) `In-degree distribution` (b) `Tweet count distribution`

Fig. 1. Distributions for social network and Twitter data

3 Tweeting and the Social Network

We begin investigating the interplay between social networking and tweeting by studying how one's tweets are related to those of his or her neighbors. More specifically, we compute pairwise similarity measures and then plot different distributions for different geodesic distances between these pairs in order to see if nodes that are closer in the social graph tend to tweet about similar topics. While this at first appears to be a straightforward experiment, there are two complicating challenges: (1) the lack of consistent and reliable tweet topic labels and (2) the complexity of pairwise similarity calculations coupled with the enormity of the social network.

First, although individual tweets lack topic labels and are often too short to reliably assign labels to, collections of user tweets can indicate coherent topics and interests. Accordingly, for each three-day interval in the tweet data, we take each user and collect all of his or her tweets into a single *document*. We choose three days as the interval length because our experiments suggest that this choice produces topics that are neither too specific nor too general.

With this, we can consider all documents in a given interval as a single collection of documents, referred to as a *corpus*. We then pass each corpus as input to David Blei's Latent Dirichlet Allocation (LDA) [2] model to assign topics to each document in that corpus. After doing this for all corpora, each user is associated with a specific time-series which indicates his or her series of topics for all 3-day intervals over the entire 7 months. Although each LDA model is built independently for each 3-day interval, which prevents us from relating topics in one interval to topics in any other interval, we can nonetheless see how a user's topic compares to other users' topics *in that interval*. Moreover, by performing an interval-by-interval comparison for a pair of users, we can potentially calculate a similarity score. Although there are likely many ways to do this, we define our pairwise similarity as follows:

$$similarity_{A,B} = \frac{\sum_{i=1}^{T} \delta(A_i, B_i)}{T}$$

where A and B are time series of document topics for two users, T is the number of intervals, and the δ function is defined as follows:

$$\delta(A_i, B_i) = \begin{cases} 1 \text{ if } A_i = B_i \\ 0 \text{ otherwise} \end{cases}$$

In other words, this score is the number of intervals for which two users tweet on the same topic, divided by all possible intervals, and is therefore in the interval $[0, 1]$.

Second, with the above similarity definition, we note that computing similarities between all nodes in the network would be $O(n^2T)$ where n is the number of nodes and T is the number of intervals. Given the enormity of the social network, this is computationally intractable. Therefore, we first only assign topic labels to documents containing at least 10 tweets and then only consider users with topic labels for at least half the intervals. This improves the reliability of our topic labeling and serves to decrease the number of nodes considered. However, since the number of nodes is still on the order of tens of thousands, we then act to decrease the computational complexity of the procedure by distributing it over thousands of machines using the software described in [6].

After taking these steps, we were then able to calculate pairwise similarities for all nodes of interest. We decomposed these similarity values based on geodesic distance and plotted similarity distributions for each distance. As can be seen in Figure 2, most pairs, irrespective of distance, tend to be dissimilar. However, we can nonetheless observe that nodes that are closer in the social graph are less likely to be highly dissimilar and generally more likely to be more similar. Thus, there seems to be some signal suggesting either tweeting influence or homophily in the social graph. Further exploration of this phenomenon might contribute to more effective viral marketing or targeted advertising on Twitter based on one's friends and followers.

4 The Retweet Network

Given the apparent correlation between social networking and tweeting, we now further explore this issue by taking advantage of twitter's "retweet" feature, which allows users to re-broadcast other users' tweets by prefixing a tweet with "RT @username." More specifically, we construct the implicit weighted retweet network where nodes are users, a directed edge from A to B exists if A retweets B's content, and all edges are weighted based on retweet frequency. This network therefore allows us to see who reads and re-broadcasts whose tweets and to mine for patterns in these behaviors. Along these lines, we now present properties and statistics of the retweet network.

First, in trying to understand how users are typically retweeted, we plot the in-degree distribution in Figure 3a. This distribution clearly demonstrates that most users are rarely retweeted while a small number are retweeted extremely often. Moreover, the entire distribution approximately follows a power law.

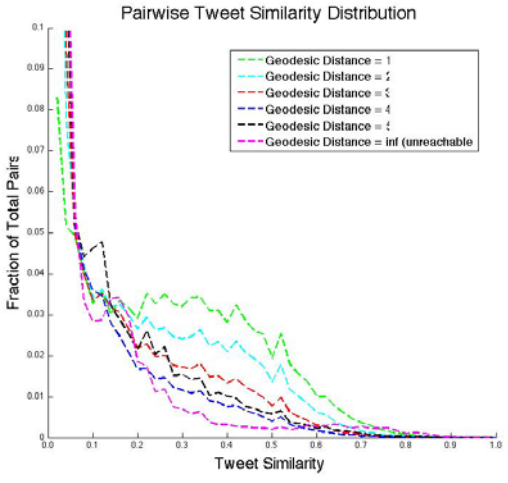
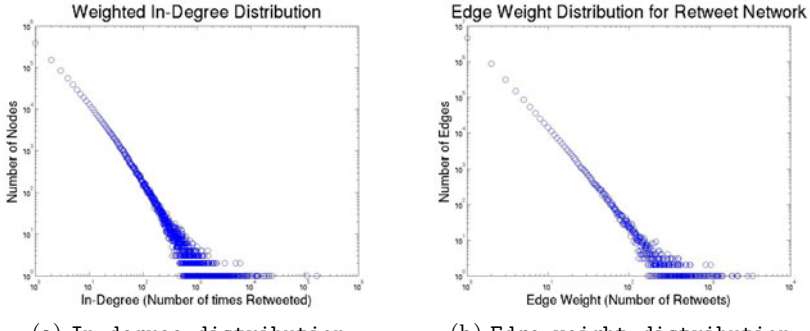


Fig. 2. Similarity distributions for various geodesic distances in the social network

Second, we also examine the nature of "retweet ties" by plotting the edge-weight distribution in Figure 3b. In doing so, we can see that the distribution very clearly fits a power law, indicating that most links exist because of a small number of retweets, but that some are highly active and represent consistent retweeting behavior.



(a) `In-degree distribution` (b) `Edge-weight distribution`

Fig. 3. Distributions for retweet network

Third, in addition to presenting distributions that pertain simply to individual nodes or edges, we also study the reciprocity of retweet edges in order to understand how these retweet behaviors are reciprocated. As can be seen by the pairwise reciprocity distribution shown in Figure 4, many relationships are highly reciprocal, but there is also a high number of non-reciprocal edges. While surprisingly high numbers of non-reciprocal edges have been observed in other

OSNs [7], the retweet network exhibits an unparalleled degree of non-recprocity on these edges, to the best of our knowledge. Indeed, since the x-axis is on log-scale, Figure 4 implies that a significant number of relationships have one user retweeting the other user's content more than 100 times more often. This phenomenon most likely stems from Twitter being more of an information network than a social network. Since the primary purpose of retweeting is to share information, certain users evidently have no problem retweeting another user's content without reciprocation, so long as it improves their ability to spread information. In light of the observed non-reciprocity, we can appreciate how any aim to disseminate information on Twitter ought to take retweet edge directions and weights into account, since the variability suggests that some edges have minimal affects while others are of great importance.

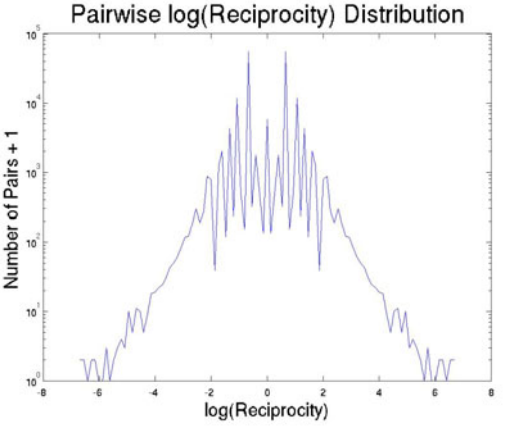


Fig. 4. Pairwise reciprocity distribution for the retweet network

Fourth and finally, we study closed triads (three fully connected nodes) and examine whether a user who appears in more of these structures is retweeted more or less frequently. To do so, we first transform any edge in the retweet digraph into an undirected edge so that the retweet network is now undirected. With this, we then perform a breadth first search at a random starting node to obtain a 10,000 node random sample, and count how many closed triads each node appears in. Then, for each unique closed triad count, we plot the average number of times retweeted for all nodes with that count. As can be seen in Figure 5, it seems that there is a clear correspondence between the number of closed triads and the number of times retweeted. However, does this mean that closed triads result in better information flow, or is this trend simply a result of users with more neighbors getting retweeted more often? In other words, are closed triads any better than open triads? Researchers have often reported that closed triads are more stable in social networks, but since Twitter is largely an information network, perhaps this is not the case here.

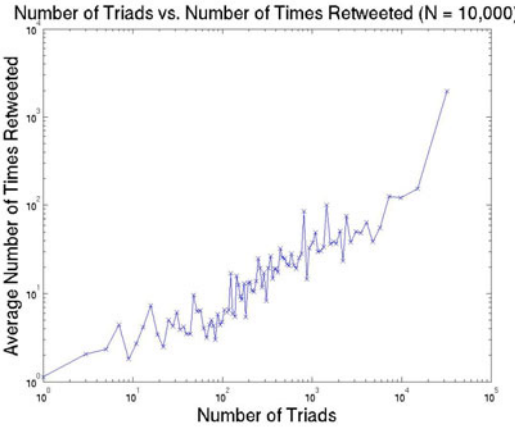


Fig. 5. Number of triads vs. average number of times retweeted for nodes appearing in that number of triads

To explore this question, we calculate the clustering coefficients for all nodes in the sample. We then bin these values, and, for each clustering coefficient bin, we plot the the average number of times retweeted for all nodes with that coefficient. In other words, we produce the same plot as is shown in Figure 5, but do so for clustering coefficient instead of closed triad count. As can be seen in Figure 6, users with lower clustering coefficients are retweeted more frequently. This suggests that in order to be a popular "retweeter," one should not actually aim to be the center of a densely connected egonet. Instead, a user should obtain as many disparate friends as possible in a sort of "star" graph structure. And indeed, the importance of these star structures is in clear agreement with the presence of "elite" Twitter users, as reported in [10].

5 URL Retweeting

With our initial understanding of the retweet network and of the interplay between local topologies and retweeting behaviors, we can now examine how these behaviors affect information spread. In order to do this, we examine the diffusion of URLs shortened on `bitly.com` for two reasons. First, tracking any URL is a reliable way to track information diffusion as each URL is unique. Second, tracking non-shortened URLs on Twitter can sometimes be misleading since we assume one information cascade when there might in fact be several different cascades originating from different users who see the URL on external sites at different times. Bitly shortened URLs, on the other hand, are more reliable as users often shorten them for the purpose of tweeting. As such, any subsequent users who retweet them have almost certainly seen them on Twitter.

In analyzing the diffusion of bitly shortened URLs, we first examine how many nodes tend to be "infected" by URL diffusion. In Figure 7, we show the

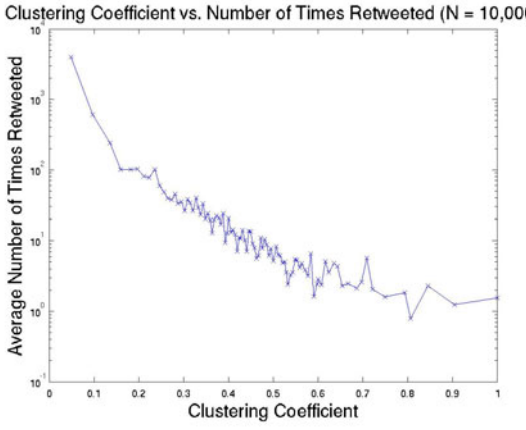


Fig. 6. Clustering coefficient (CC) vs. average number of times retweeted for nodes with that CC

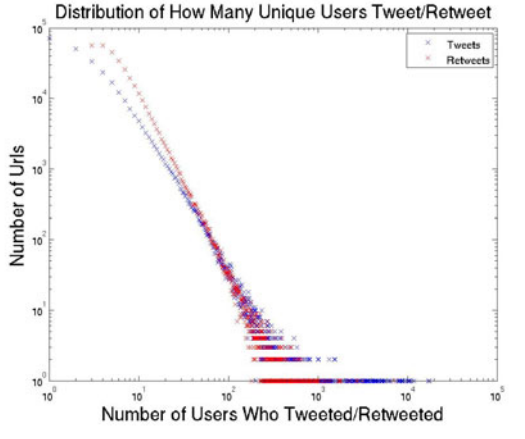
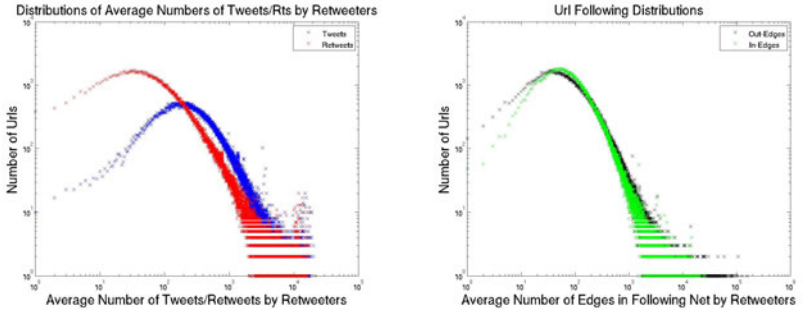


Fig. 7. Average number of users who tweeted / retweeted vs. number of URLs with that number of users

distribution of infected population sizes, which fits a clear power-law except for the head of the distribution, which indicates that a less than expected number of URLs infect very few users. We also inspect the kinds of nodes that retweet urls. In Figure 8a, we illustrate the average number of tweets or retweets by users who retweet URLs, and plot the distribution of these counts. As can be seen, most URLs are retweeted by users with tens or hundreds of tweets/retweets, but a small number of URLs are retweeted by extremely active users. Moreover, in Figure 8b, we illustrate the average number of social ties by users who retweet URLs and observe a similar trend.



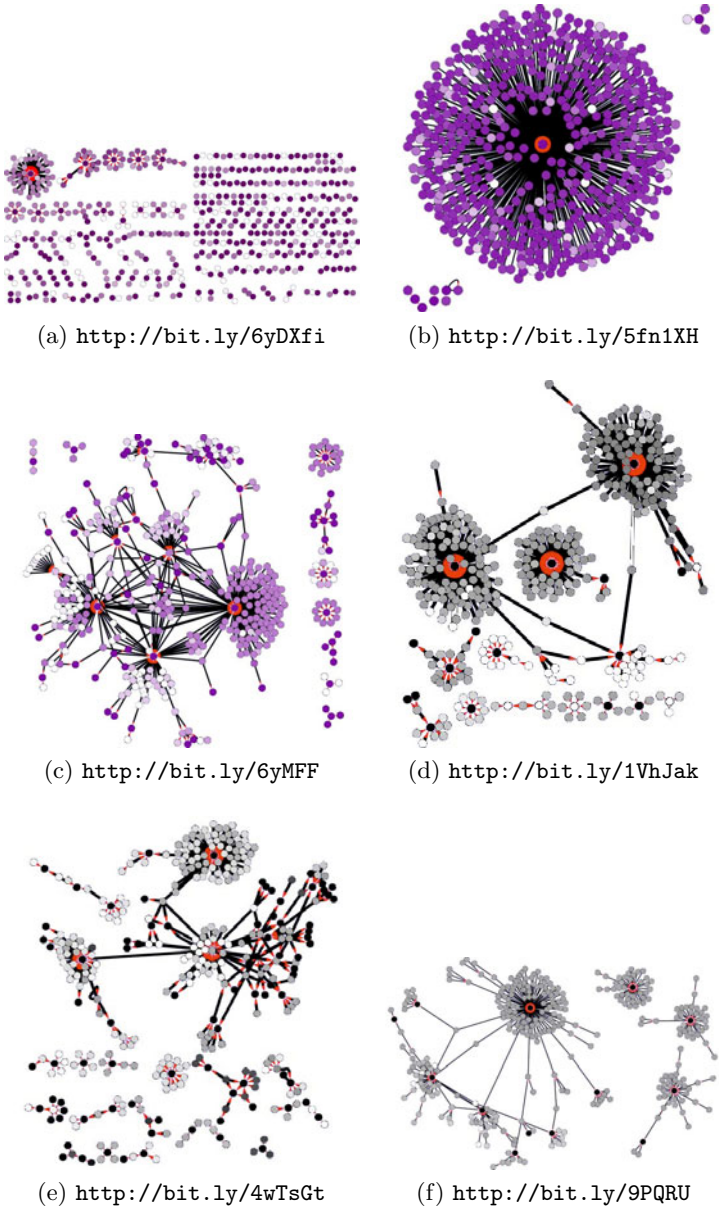
(a) `Average tweets/retweets frequencies.` (b) `Average in/out degree.`

Fig. 8. URL distributions

Finally, in addition to simply looking at distributions, we also examine the diffusion of individual URLs. In Figure 9, we visualize the diffusion of six different bitly shortened URLs. Each visualization shows a retweet subgraph where nodes are users and a directed edge goes from A to B if A retweets $B's$ tweet that included a reference to that URL. The node colorings represent time of infection, where darker colors retweeted the URL earlier than light colors. In looking at these visualizations, we can recognize that diffusion generally originates in one or more "hub" nodes that are frequently retweeted. In light of the clustering coefficient distribution shown earlier, we can better understand how users who are retweeted often generally have sparsely connected egonets. After all, most of the hubs in these URL retweet networks are connected to nodes that are not connected to one another.

6 The Social and Retweet Networks

Now that we have presented our analyses of the explicit social network and the implicit retweet network, we are well prepared to consider how these two networks are related. In particular, since the same nodes appear in both networks, we can explore how a node's topological characteristics in one network corresponds to its characteristics in the other. We specifically address this issue by plotting the social network in-degree versus the average retweet in-degree (Figure 10a) and versus the average retweet out-degree (Figure 10b). In looking at these plots, we can see that a high in-degree in the social network generally corresponds to high activity in the retweet network (many tweets and many times retweeted). This makes intuitive sense as one would assume that having many followers would generally indicate an active Twitter user who might both retweet and be retweeted.

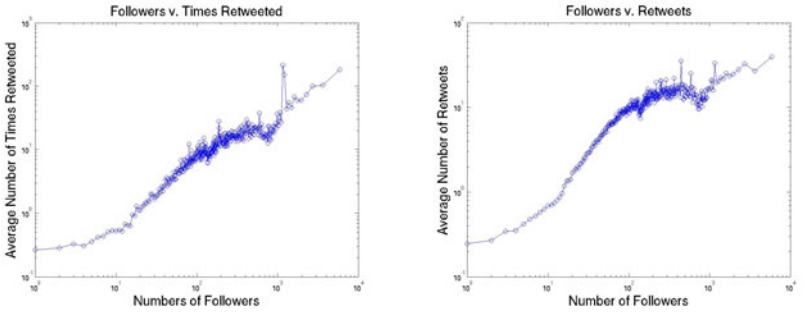


(a) http://bit.ly/6yDXfi (b) http://bit.ly/5fn1XH

(c) http://bit.ly/6yMFF (d) http://bit.ly/1VhJak

(e) http://bit.ly/4wTsGt (f) http://bit.ly/9PQRU

Fig. 9. Diffusion of URLs in retweet networks. Node coloring indicates infection time (darker for earlier infection times). Edges are directed with red arrows at target nodes.



(a) (b)

Fig. 10. Plots of social network degree vs. average retweet network degree for all nodes with that social network degree

7 Conclusion and Future Work

In this paper, we presented analyses of the explicit social network and the implicit retweet network, focusing on macro properties and local graph structures, in order to better understand the relationship between socialization and tweeting. We first utilized a topic model and a corresponding similarity measure and showed that users who are closer in the social network tend to be slightly more similar. We then analyzed the retweet network and found that many edges are highly non-reciprocal and that frequently retweeted nodes often have low local clustering coefficients. Next, we analyzed the diffusion of individual URLs, and our visualizations of the diffusion networks supported the observed low clustering coefficients. Finally, we concluded by showing that a node's in-degree in the social network generally corresponds to its retweeting activity.

While these findings shed light on various Twitter features and phenomena, they also indicate that these issues require further exploration. For example, although we found some evidence that tweet topic similarity correlates with proximity in the social graph, we could better understand this interplay if we had a more effective method to assign topics. In particular, rather than binning tweets and using a topic model, it might be interesting to explore methods from natural language processing that can assign topics to individual tweets. In addition, after observing low clustering coefficients for nodes that are frequently retweeted, it might be interesting to further explore what kinds of graph structures these nodes appear in, using graphlet and/or motif analyses. The temporal data might also be further leveraged and used to conduct a more thorough temporal analyses of the networks. Finally, given this work, we might aim to develop machine learning algorithms that can infer missing information about the the social network, the retweet network, or tweeting behaviors by incorporating knowledge from all three.

Acknowledgments. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 and in part by the National Science Foundation Grant BCS-0826958. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

We would like to thank Haewoon Kwak at KAIST, as well as Jure Leskovec at Stanford University, for providing the data used in the study.

References

1. Bakshy, E., Hofman, J., Mason, W., Watts, D.: Identifying influencers on twitter. In: Fourth ACM International Conference on Web Seach and Data Mining (WSDM)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
3. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 675–684. ACM, New York (2011)
4. Dodds, P., Harris, K., Kloumann, I., Bliss, C., Danforth, C.: Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. Arxiv preprint arXiv:1101.5120 (2011)
5. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, pp. 591–600. ACM, New York (2010)
6. Lichtenwalter, R., Chawla, N.: DisNet: A Framework for Distributed Graph Computation
7. Lussier, J., Raeder, T., Chawla, N.: User generated content consumption and social networking in knowledge-sharing osns. In: Chai, S.-K., Salerno, J.J., Mabry, P.L. (eds.) SBP 2010. LNCS, vol. 6007, pp. 228–237. Springer, Heidelberg (2010)
8. Romero, D., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 695–704. ACM, New York (2011)
9. Welch, M., Schonfeld, U., He, D., Cho, J.: Topical semantics of twitter links. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 327–336. ACM, New York (2011)
10. Wu, S., Hofman, J., Mason, W., Watts, D.: Who says what to whom on twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 705–714. ACM, New York (2011)