Exploring Twitter Networks in Parallel Computing Environments

Bo Xu
Northeastern University, China
No.11, Lane 3, Wenhua Road
Shenyang, Liaoning, China
xubosuper@163.com

Yun Huang, Noshir Contractor
Northwestern University
2145 Sheridan RD, TECH C210
Evanston, IL 60208
{yun,nosh}@northwestern.edu

ABSTRACT
Millions of users follow each other on Twitter and form a large and complex network. The size of the network creates statistical and computational challenges on exploring and examining individual behavior on Twitter. Using a sample of 697,628 Korean Twitter users and 34 million relations, this study investigates the patterns of unfollow behavior on Twitter, i.e. people removing others from their Twitter follow lists. We use Exponential Random Graph Models (p*/ERGMs) and Statnet in R to examine the impacts of reciprocity, status, embeddedness, homophily, and informativeness on tie dissolution. We perform data processing, statistics calculation, network sampling, and Markov chain Monte Carlo (MCMC) simulation on Gordon, a unique supercomputer at the San Diego Supercomputer Center (SDSC). The process demonstrates the role of advanced computing technologies in social science studies.

Categories and Subject Descriptors
H5.0. Information interfaces and presentation: General.

General Terms
Management, Measurement, Performance, Theory.

Keywords
Twitter, Social network analysis, Parallel computing, Exponential Random Graph Model, ERGM

1. INTRODUCTION
The rapid development of online social media Twitter has fundamentally changed people's way of communication and getting information. Twitter users interact by posting tweets, sharing photos and sending private messages. With the widely adoption of its services all over the world, Twitter now is not only a social platform for ordinary user communication but also a broadcasting channel for commercial users' advertising and marketing. So it is of vital importance to understand Twitter users' behavioral patterns and provide insightful suggestions on further development of Twitter services for both communication and commercial uses.

The primary purpose of this research is to provide a complete picture of the Twitter network. We try to calculate its network structural properties, identify different Twitter user groups and detect local components within Twitter. We also study the removal of follow relations. Tie formation and dissolution are the two fundamental factors that drive the structural evolution of a network. Numerous studies focus on the intrinsic nature of tie formation while little work has been done to explore tie dissolution, though they are equally important. In this paper, we analyze the phenomenon of tie dissolution on a longitudinal dataset from Twitter, i.e. unfollow behavior on Twitter. Twitter users may subscribe to others' tweets, known as *following*, and become followers of other users. Users are free to unsubscribe and remove others from their following lists, known as *unfollowing* [6,7].

Because of the interdependency of unfollow relations, we use an Exponential Random Graph Model (p*/ERGM)[10]to analyze the likelihood of unfollow relations. ERGM is a statistical model which examines whether specific network structures can influence the emergence of a tie given the attributes of users and the relations among them in a network. Since ERGM can only explain what influences the formation of a tie, here in this paper, we create an "*unfollow network*" to study the breaking of ties. Based on samples at two time points in a Twitter network, we consider a directed link from A to B indicating the *unfollow relation* between the two if A followed B at time 1 but stopped the following at time 2. In this way, the dissolution of following relations is modeled by the emergence of an unfollow tie in the unfollow network.

All the analysis are based on two snapshots of 697,628 Korean Twitter users and 34 million following relations among them. Since the scale of the dataset is very large, we utilize Gordon [12], a unique supercomputer at the San Diego Supercomputer Center (SDSC), to process the dataset and perform statistical and empirical analyses: Python is used for parsing the raw data files and construct user attributes and relation networks, the igraph package in Python is used to calculate network statistics, and Statnet in R is used to estimate p*/ERGM models using Markov Chain Monte Carlo (MCMC) simulation.

2. DATA AND HYPOTHESES

XSEDE '13, July 22 - 25 2013, San Diego, CA, USA

Since cultural beliefs about relationships may vary, we focus on a set of users in the same cultural context. Based on a sample of

Table 1. Statistics of the follow and unfollow networks (both directed)

	Follow network at time 1	**Unfollow relations at time 2**
Vertices	697,628	211,263 involved
Arcs	34,429,170	858,702 unfollow
Mutual pairs	12,676,988	33,015 reciprocal unfollow

*697,628*Korean Twitter users, we take two snapshots of their follow networks on June 25th, 2010 (as time 1) and on April 26th, 2011(as time 2). By comparing the two snapshots, we detect removed follow relations and construct an unfollow relation network. There were34 million follow relations at time 1, 73.6% of which are mutual. At time 2, 7.7% of the following relations disappeared. Table 1 shows the detailed statistics.

Degree distribution is a statistic that measures a networks' structural property. It specifies the probability P(k) that a randomly selected vertex has k edges. Since the network we constructed is a directed network, we calculate both in-degree and out-degree distributions to illustrate the global property of the given Korean Twitter network (see Figure 1).

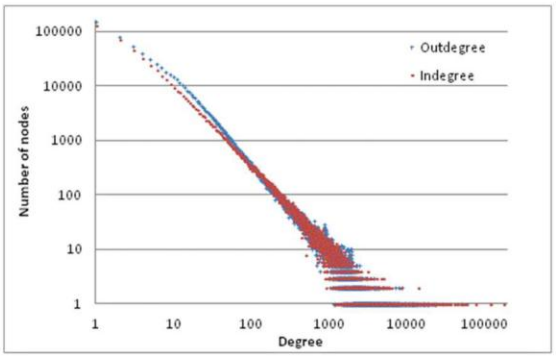

Figure 1. Indegree and outdegree distributions of the follow network on June 25th 2010

The classical complex network theory asserts that the degree distribution of most networks in reality follows a power-law distribution, which is a straight line in the log-log plot. However, Figure 1 shows clearly that fat tails in both in-degree and out-degree distributions contradict the power-law assumption. Figure 2 shows the joint distribution of indegree and outdegree for users with less than 2,000 links. The degrees of these users are consistent with the power-law distribution. Most of users only follow and are followed by less than 200 others and a few users have more links. Some users with more than a few thousands of links form the fat tails. In Twitter, most users with more 2,000 followers are popular hubs and whereas users with more than 2,000 followees usually take a special effect to boost their follower population.

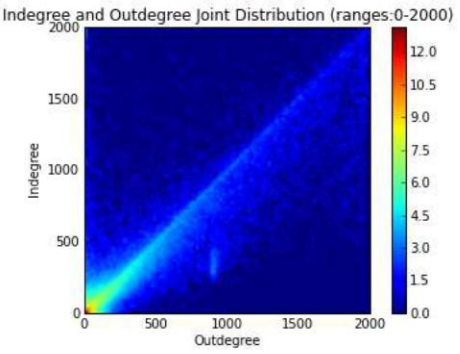

Figure 2. The heat map of indegree and outdegree joint distribution user groups (color indicates the number of users in log-scale)

2.1. Identification of Twitter User Groups

Basic on the statistical analysis, we found that the Twitter network is composed of 3 types of users: ordinary users (e.g. friends in communities), commercial users (e.g. interactive marketers and small business), and super stars (e.g. celebrities, athletics, and media). Super stars usually have tens of thousands followers but only follow a few others (i.e. extremely high in-degrees but low out-degree/in-degree ratios). Users with a large amount of mutual followers are commercial and they follow many other people only expecting them to follow back in return. Many of them, so called interactive marketers, use their connections to promote other accounts or messages as a business in Twitter.

Apparently, the behavioral patterns of these three types of users are different. It is important to identify the different user groups, and study their behaviors respectively. Here, we cluster the three user groups based upon users' in-degree, out-degree and in-degree/out-degree ratio. We use the K-means algorithm to perform the analysis on Gordon. The three types of users can be illustrated in Figure 3.

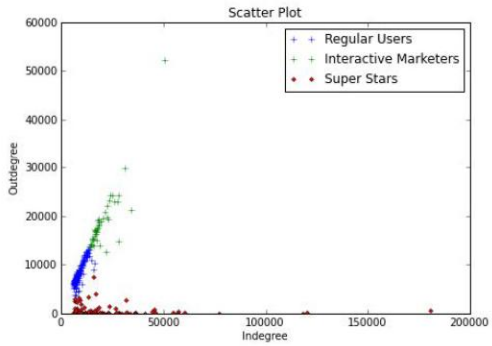

Figure 3. An illustration of three types of Korean Twitter user groups

Taking 2,000 as the outdegree cut off point, the dataset contains 695,978 regular users and 1,650 super stars or interactive marketers. Taking the cut-off point of indegree/outdegree ratio at 1.8, we classify 17 users as super stars and 1,633 as interactive marketers. To test the outcome of our analysis, we checked the personal status of each user on Twitter. More than 95% of the users are correctly grouped by the algorithm.

2.2. Local samples in Korean Twitter network

To investigate the local properties of the network, we also analyze the local components in our research. The whole network contains 162,363 components. The largest connected component (LCC) has 526,944 nodes, the second LCC only has little more than 100 nodes. The network has more than 100,000 isolated pairs of users. This statistic shows that most vertices within Twitter network are tightly connected. The component size distribution can be illustrated as follows:

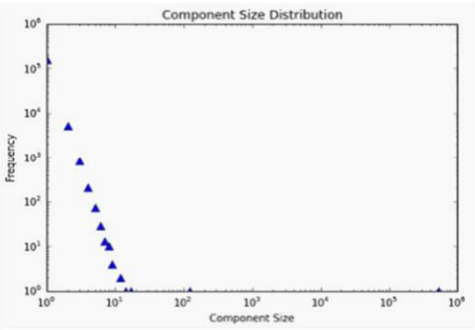

Figure 4. Component size distribution of the Korean Twitter network.

Since the computation complexity of MCMC simulation on random graphs is NP hard, R/statnet package cannot estimate a network with more than a few thousands of nodes. With the help of multiple computational units and fast flash based storage system, we design a sampling approach to extract local networks and examine each sample using independent computer jobs.

As we discussed above, different user groups exhibit different user behaviors. The motivations of unfollow in ordinary user groups and commercial user groups also vary from each other. Our research only focuses on ordinary users to explore the factors that may influence their unfollow behavior. We use one-wave snowball sampling to extract closely connected communities of ordinary users. The sampling method can be described as follows: First, select a seed user and find all his/her followers at time 1; second, select those involved in at least one unfollow relation at time 2; and third, remove the top 2.5% outliers with extremely large in-degree or out-degree in the unfollow network. This approach breaks the huge Twitter network into many small communities and focuses on the unfollow activities at a more normalized level.

394 snowball samples are generated using random seeds with 1,000 to 2,000 followers to capture communities related to popular but not commercial twitter users. Each sample is estimated independently and all results are combined in a macro-level between-community analysis by using meta-analysis [11].

2.3. Hypotheses about Unfollow on Twitter

Previous studies on Twitter show that *reciprocity* [2,3], *embeddedness* [4,15], *social status* [1], *homophily* [9] and *informativeness* [8] play key roles in the process of tie formation. We adopt the set of theoretical frameworks from previous studies of tie formation to the context of tie dissolution on Twitter and explore their impacts with the following six hypotheses.

- Hypothesis 1: Users are less likely to unfollow those who follow them.
- Hypothesis 2: Users are more likely to unfollow those who unfollow them.
- Hypothesis 3.1: Users with more followers are more (less) likely to unfollow (be unfollowed).
- Hypothesis 3.2: Users with more followees are less (more) likely to unfollow (be unfollowed).
- Hypothesis 4: Users are less likely to unfollow those with whom they share more common followees.
- Hypothesis 5: Users are less likely to unfollow those who are interested in similar topics.
- Hypothesis 6: Users are less likely to unfollow those whom they have retweeted, mentioned, replied, or favorited.

RESULTS & DISCUSSION

As in a logistic regression, an ERG model characterizes the impact of explanatory variables on the log odds of unfollow relations. A positive estimated parameter indicates that a larger-valued corresponding explanatory variable leads to a higher tie probability, conditional on all other effects in the model. Table 2 gives a summary of our model results. Other than the 10 explanatory variables drawn from the hypotheses discussed above, the structural variable *Edges, Weighted # out-stars* and *Weighted # in-stars* are also included to control for the network structure. In addition, the follow network at time 1 is used as a base line to prevent the generation of unfollow relations between users who are not connected at time 1 during the Monte Carlo Markov Chain Simulation in ERGM [5]. R/statnet package 3.0-1 is used for the ERGM analyses.

The results show that the reciprocity plays a critical role both in follow relations and unfollow relations. When two users follow each other mutually, the odds ratio of one unfollowing another is only 7.9%, i.e. exp(-2.54), of those without mutual following relations. Mutual ties indeed make the relation stronger and more cohesive. However, if one user in a mutual relation unfollows the other, the unfollowed user is very likely to unfollow in return. The odds ratio of reciprocal unfollowing is 2.5 times, i.e. exp(3.45-2.54), of that of initial unfollowing. Both Hypotheses 1 and 2 are supported.

The opposite impacts of the number of followers and the number of followees reveal the difference in social status, especially among senders, who initiate unfollow actions. High status users, who have more followers and less followees, are more likely to unfollow others and less likely to be unfollowed.

As predicted in Hypothesis 4, the number of *common followees* has a negative impact on unfollow relations. Users connected by many common friends are embedded in a strong network structure and therefore have more persistent ties. On the other hand, the number of *common hashtags* has no significant impact. There is no evidence that the common interests reduce the likelihood of unfollowing. The homophily effect proposed in Hypothesis 5 is not supported.

Contrary to Kwak et al. [8], we do not find any significant impacts of *replies*, *retweets*, *mentions,* and *favorites* on unfollow behavior. Since our samples focus on small and tightly connected communities, users interact for relational rather than informational purposes. The frequency of their interactions may not affect the persistence of their social ties.

Table 2. Summary of model results.

Parameters	Estimate		
Unfollow network structures:			
Reciprocity	3.45*	H2	Supported
Sender attributes of unfollow relations:			
# Followers	0.21*	H3.1	Supported
# Followees	-0.47*	H3.2	Supported
Receiver attributes of unfollow relations:			
# Followers	-0.03*	H3.1	Supported
# Followees	0.07*	H3.2	Supported
Community in follow networks at time 1:			
Mutual ties	-2.54*	H1	Supported
Common followees	-1.83*	H4	Supported
Common Hashtag	-0.03	H5	Not supported
Sender's interactions to receiver:			
# Replies	0.001	H6	Not supported
# Retweets	-0.009	H6	Not supported
# Mentions	0.321	H6	Not supported
# Favorites	-0.003	H6	Not supported
Unfollow network structures as control variables:			
Edges/density	-1.30*		
Weighted out-stars	-0.57*		
Weighted in-stars	0.19*		

Note: * indicates p<0.001, # followees, # followers, and # replies are in thousands.

Compared to logistic regression models, the ERG models consider important network structures when examining the impacts of explanatory variables. The number of edges, like the constant (or intercept) in logistic regressions, controls for the network density; the weighted number of out-stars controls for the out-degree distribution, i.e. the overall tendency of people unfollowing others; the weighted number of in-stars controls for the in-degree distribution, i.e. the overall tendency of people being unfollowed. The negative coefficient of *edges* suggests that the density of the unfollow networks is very low and people are less likely to unfollow others randomly. In fact only 2.5% of follow relations have dissolved in ten months. The negative coefficient of *weighted out-stars* indicates that there are more users with high out-degrees than expected and they unfollow many others. On the other hand, the positive coefficient of *weighted in-stars* indicates that users are less likely to be unfollowed by many others at the same time.

The parallel computing facility greatly improved the process of building empirical models for analysis. The MCMC simulation for estimating ERG models is inherently a serial algorithm. Figure 5 shows the run time for 100 sample networks using Statnet in R. The estimation runs took 5 to 35 minutes each according to the sizes of the networks. The total run time of all models is around 24 hours if we run it on a single computer. However, using multiple cores, it took only 46 minutes as jobs in the normal queue and speeds up 30 times on Gordon. This makes it possible to explore different models interactively and expands our ability to explore complicated behavior in Twitter networks. Furthermore, we are expect an almost linear performance gain when more samples are tested simultaneously.

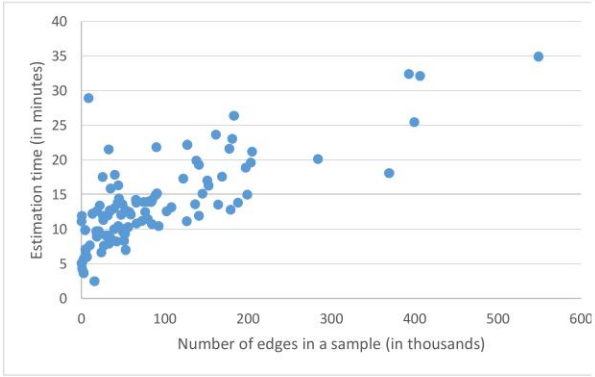

Figure 5. Numbers of edges of 100 network models and their estimation time using Statnet.

CONCLUSION

This paper explores a large Twitter network in a parallel computing environment. Gordon supercomputer provides computational power to process a big network file, calculate complex network statistics, and estimate statistical models.

Using ERG models, we show that unfollow links are interdependent: they are highly reciprocal and clustered. The breakup of one tie in a pair of mutual follow relations will lead to the breakup of the other. Some users tend to unfollow many others. We also find that embeddedness and social status reflected by network structures have a strong impact on unfollow decisions in Twitter. Our study finds no significant impact of homophily and informativeness on tie breaking in our samples. This suggests that the use of Twitter in small communities is more relational rather than informational.

Our samples focus on small and tightly connected communities and therefore reveal more relational factors in tie dissolution. In other types of communities, such as a group of followers of a celebrity, information oriented factors such as homophily based on common interests and informativeness may have a greater impact on unfollow behavior. Future research, should evaluate unfollow patterns in Twitter communities of different types and sizes.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation (Grant No. CNS-1010904, OCI-0904356, OCI-1053575, & IIS-0841583) and Army Research Lab (W911NF-09-02-0053), the National Natural Science Foundation of China (No. 90924020), MOE project of Humanities and Social Sciences (No. 11YJA630044), and the Innovation Foundation of BUAA for PhD Graduates. We thank Sue Moon for providing the data sets.

REFERENCES

[1] Emerson, R. M., Power-Dependence Relations, *American Sociological Review,* 27, 1 (1962), 31-41.

[2] Golder S.A. and Yardi, S., Structural Predictors of Tie Formation in Twitter: Transitivity and Mutuality, in *IEEE Social Computing* (2010), 88-95.

[3] Gouldner, A. W., The Norm of Reciprocity: A Preliminary Statement, *American Sociological Review,* 25, 2 (1960), 161-178.

[4] Gruzd, A., Wellman, B., and Takhteyev, Y., Imagining Twitter as an Imagined Community, *American Behavioral Scientist,* 55, 10 (2011), 1294-1318.

[5] Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., and Morris, M., ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks, *Journal of Statistical Software,* 24, i03 (2008).

[6] Kivran-Swaine, F., Govindan, P., and Naaman, M., The impact of network structure on breaking ties in online social networks: unfollowing on twitter, in *Proc. CHI 2011*, ACM Press (2011), 1101-1104.

[7] Kwak, H., Chun, H., and Moon, S., Fragile Online Relationship: A First Look at Unfollow Dynamics in Twitter, in *Proc. CHI 2011*, ACM Press (2011), 1091-1100.

[8] Kwak, H., Moon, S., and Lee, W., More of a receiver than a giver: Why do people unfollow in Twitter, in *Proc. ICWSM 2012* (2012).

[9] McPherson, M., Smith-Lovin, L., and Cook, J. M., Birds of a Feather: Homophily in Social Networks, *Annual Review of Sociology,* 27, (2001), 415-444.

[10] Robins, G., Pattison, P., Kalish, Y., and Lusher, D., An introduction to exponential random graph (p*) models for social networks, *Social Networks,* 29, 2 (2007), 173-191.

[11] Snijders, T. A. B. and Baerveldt, C., A multilevel network study of the effects of delinquent behavior on friendship evolution, *The Journal of Mathematical Sociology,* 27, 2-3 (2003), 123-151.

[12] Strande, S. M. et al., Gordon: Design, Performance, and Experiences Deploying and Supporting a Data Intensive Supercomputer, in *Proc XSEDE12, ACM press (2012).*