How Social Network is Evolving?
- A Preliminary Study on Billion-scale Twitter Network

Masaru Watanabe
Tokyo Institute of Technology and JST CREST
2-12-1 W8-E, O-okayama, Meguro-ku,
Tokyo, 152-8550, Japan
watanabe.m.ay@m.titech.ac.jp

Toyotaro Suzumura
Tokyo Institute of Technology, IBM Research Tokyo
and JST CREST
2-12-1 W8-E, O-okayama, Meguro-ku,
Tokyo, 152-8550, Japan
suzumura@cs.titech.ac.jp

ABSTRACT

Recently, social network services such as Twitter, Facebook, MySpace, LinkedIn have been remarkably growing. There are various studies about social networks analysis. Haewoon performed the analysis of the Twitter network on 2009 and shows the degree of separation. However, the number of users on 2009 is about 41.7 million, the graph scale is not very large compared with the current graph. In this paper, we conduct a Twitter network analysis in terms growth by region, scale-free, reciprocity, degree of separation and diameter using Twitter user data with 469.9 million users and 28.7 billion relationships. We report that the value of degree of separation is 4.59 in current Twitter network through our experiments.

Categories and Subject Descriptors

G.2.3 [**Discrete Mathematics**]: Applications

Keywords

Twitter, Social Network Analysis, Degrees of Separation, Diameter, Reciprocity, Degree Distribution

1. INTRODUCTION

Recently, social network services such as Twitter, Facebook, MySpace, LinkedIn have been remarkably growing. These services take part in as mediums that support a connection between people. A lot of people use these social tools and it is expected that the number of users increase further in the future.

We can regard the connections between people on the social services as a graph structure. However, each graph on the social services has a different feature. For example, on Facebook, a user creates an account with his true name and makes some friends. So the connections of people on Facebook have close property to human relationship in the real world. On Twitter, a user can easily get information by following his interesting person. So the connections of people on Twitter consist of the interests of people. MySpace centers on music and entertainment, so the graph on MySpace are structured based on community of people having a same interest.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

There are various studies that analyze the social networks. Haewoon [4] performed the analysis of the Twitter network on June 2009 and showed the degree of separation. However, the number of users on 2009 is about 41.7 million, the graph scale is not very large compared with the current graph. Figure 1 shows the transition of the number of users on Twitter from June 2006 to September 2012. The number of users on September 2012 attains 469.9 million and the number of relationships attains 28.7 billion. This data collection is obtained by our series of crawling for 3 months conducted in late 2012. Therefore, it is considered that with increasing users, the graph characteristics has changed greatly and we analyzed for the current large graph. This is the motivation on how such characteristics is changed and evolving from the results in 2009.

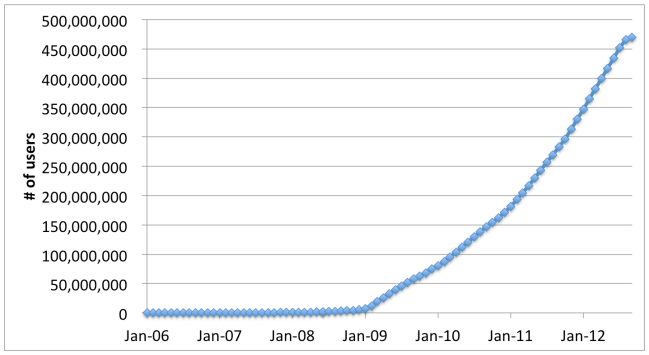

Figure 1. The transition on the number of Twitter users from 2006 to 2012

Lars Backstrom [5] computed degree of separation on Facebook social network using graph analysis tool, HyperANF [6]. The graph on Facebook seems to have a similar graph structure to the real world relationships, so the result of experiment is 4.74, which is smaller than they expected and becomes a hot topic in the world. But if we apply the "degree of separation" calculation to the Twitter network, what kind of results can we obtain? As previously mentioned, Twitter and Facebook have a different graph structure, thus it is significant to analyze the current Twitter network against Facebook.

This paper is organized as follows. Section 2 describes Twitter service overview and user data that we have collected for tree months. We conduct basic analysis of the growth of the Twitter network in terms of locality in Section 3. In Section 4 we conduct Twitter network analysis for degree distribution, reciprocity, degree of separation and diameter. Finally, in Section 5 we conclude.

2. TWITTER as Large-Scale Social Network

2.1 Twitter Service

Twitter is an online networking service and micro-blogging service on which its users can contribute short text-based messages of up to 140 characters called "tweet". Tweets are pushed onto their own stacks called "timeline", and every user can read another user's timeline. Users can display other user's tweets on their own timeline by following the one. On Twitter, the following direction becomes a directed edge between two users and forms a large-scale social graph over time.

2.2 Edge direction

Here, we briefly explain the graph structure of Twitter network. An edge on twitter network is generated when some user follows another. Figure 2 shows the graph structure, when user A follows user B. Note that, the direction of the edge is equivalent to the direction of information flow. At this time, B is a friend of A and A is a follower of B. On Facebook, graph data is undirected graph because it is necessary to obtain adversary approval to make relationship. But on Twitter, the graph structure is directed graph, because everyone can follow someone freely.

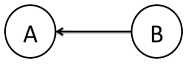

Figure 2. Edge direction when A following B

2.3 Crawling Large-Scale Twitter Data

We crawled and collected data of 469 million users that consists of two kinds of information on Twitter from July 2012 to October 2012 using Twitter API. One is user profile information in serialized XML format that contains user id, user name, user brief description, account creation time, time zone, the number of tweets and so on. The other is follower-friend information in CSV format as shown in Figure 3. In fact, we analyzed Twitter network with edge lists generated from "user id" and "follower id list" in the follower-friend data. Note that we refer to edge lists for experiment in Section 5.

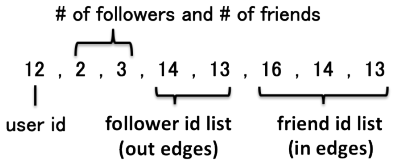

Figure 3. Example of follower-friend data in the CSV format

In order to collect the data, we began with top 1,000 users with the largest number of followers and crawled breadth-first along the direction of followers. We have collected user data for three months and quit the crawling at the end of the search 29th, though we have not collected all user data. Because the number of users that we collected after the search 26th was less than 100, so it was difficult to collect more user data. Finally, we collected 469 million user data. Serialized user profile and compressed follower-friend data size are 91GB, 231GB, respectively.

3. BASIC ANALYSIS

In this section, we describe the basic analysis of Twitter network using user profile data that we have collected for three months.

3.1 Monthly increase of users

First, we investigated how the Twitter network was grown from the viewpoint of the number of users. In conducting the investigation, we used Apache Hadoop [9] as an analysis tool and accumulated monthly increase of users from June 2006 to September 2012. Figure 4 shows the result of accumulation.

Looking at Figure 4, it can be seen that the number of users on Twitter increase explosively at the beginning of 2009 and the increase of users per month is more than 16 million people in 2012 except for October. The reason of the decline at October 2012 is that we stopped data crawling at the middle of October. However, the increase of users is not in monotonically increasing throughout. In order to clarify the cause we conducted further analysis that takes into account the locality, so we described the result in Section 3.2.

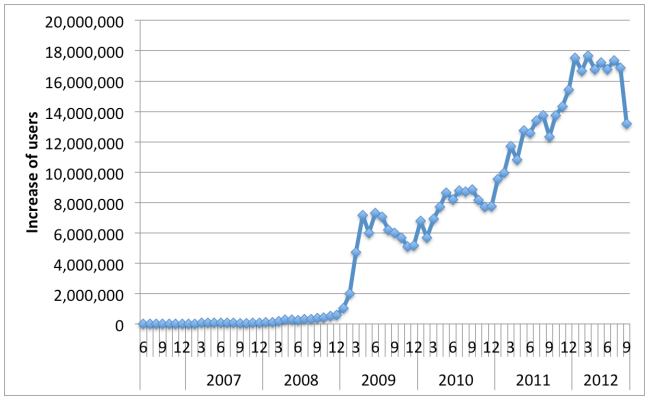

Figure 4. Monthly increase of users

3.2 Increase of users by region

In this section, we briefly mention where, when and how much growth Twitter service. We used location information in user profile data and investigated the growth of Twitter in order to obtain more detailed results. User profile data contains "location" and "time zone" properties as location information. We would like to have used "location" property to specify where users access Twitter service as accurately as possible, but it was difficult to do that because the "location" property allows users to write freely. So we used "time zone" property at the expense of a little accuracy. Users can choose about 150 candidates of cities or areas such as "Central Time (US & Canada)", "Quito", "Brasilia", "Santiago", and so on. Therefore, we associated manually the candidates of "time zone" property with countries, geographical sub-regions and regions based on United Nations Statistics Division [1]. However there are a lot of users whose "time zone" property is still not set (default: "null" or "Hawaii"), so we accumulated without such users. At this time, the number of users that were not excluded from the accumulation was about 131 million out of 469 million users.

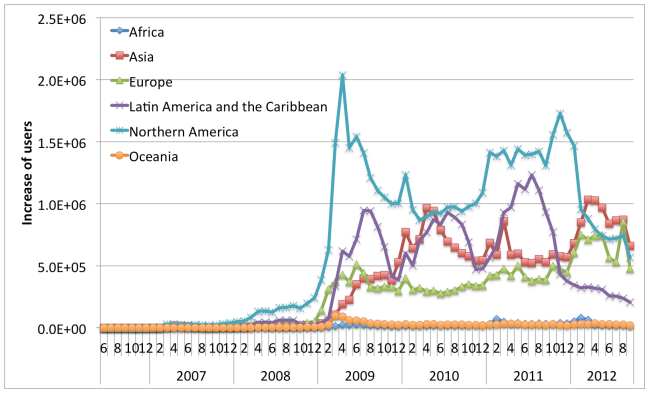

Figure 5. Transition of increase of users by region

Figure 5 shows transition of increase of Twitter users by region. As shown in Figure 5, Twitter services rapidly prevailed in Northern America from the beginning of 2009, followed by Latin America and the Caribbean, Asia and Europe.

	July 2009		October 2012	
Africa	132,476	0.66%	1,276,914	0.96%
Asia	1,659,319	8.30%	27,441,905	20.82%
Europe	3,012,827	15.07%	19,840,979	15.05%
Latin	3,802,882	19.02%	28,528,793	21.64%
NA	10,922,270	54.64%	53,179,750	40.35%
Oceania	458,284	2.29%	1,520,440	1.15%
Total	19,988,058	100%	131,788,781	100%

Table 1. The number of users in Africa, Asia, Europe, Latin America and Caribbean (Latin), Northern America (NA) and Oceania on July 2009 and October 2012

Haewoon's study with Twitter network was performed in 2009, when the number of users greatly increased, but more than 73% of users access Twitter from Americas (Northern America or Latin America and the Caribbean). On the other hand, the current Twitter network in 2012 becomes far larger than 2009 and seems to construct more complicated network because the number of user increase not only in Americas but also in various regions, especially in Asia. So it is significant to perform a new analysis and compare the results.

4. NETWORK ANALYSIS

In this section, we describe the basic analysis of Twitter network using user profile data and follower-friend data.

4.1 Degree Distribution

"Scale-free" is one of the features of a social graph. A scale-free network is a network whose degree distribution follows a power law, at least asymptotically. In scale-free network, some nodes are connected with a lot of other nodes and have a large degree, on the other hand, majority of nodes are not connected only with a very few nodes and have a small degree.

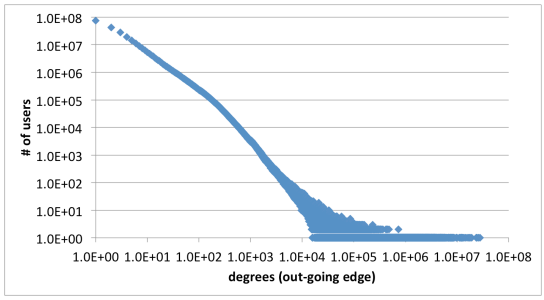

Figure 6. Degree Distribution of Followers

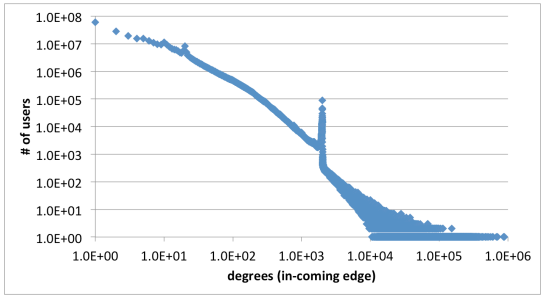

Figure 7. Degree Distribution of Friends

Figure 6 and 7 display the degree distribution of follower (out-edge) and friend (in-edge), respectively. As seen from the figures, the degree of a part of users is very large value while the degree of majority of users is small value.

We briefly explain that there are glitches at x = 20 and 2000. These glitches ware reported in [4] [8] in details. The first glitch is caused by a service on Twitter that recommends an initial set of 20 people a newcomer can follow by a single click and quite a few people take up on the offer. The second glitch is caused by upper limitation of the number of friend before 2009. Twitter removed this cap and there is no limit now.

4.2 Reciprocity

Reciprocity is a quantity to specifically characterize directed networks. We used traditional definition of reciprocity as follows:

$$r = \frac{L^{\leftrightarrow}}{L}$$

$L^{\leftrightarrow}$: The number of edges pointing in both directions

L : The total number of edges

With this definition, $r = 1$ is for a purely bidirectional network while $r = 0$ is for a purely unidirectional one. Basically, real networks have an intermediate value between 0 and 1. Table 2 displays the comparison Twitter network on 2009 and 2012.

	July 2009	September 2012
# of users	41.6 M	465.7 M
# of edges	1.47 B	28.7 B
Reciprocity	22.1 %	19.5%

Table 2. Comparison of reciprocity

In Section 3.2, we described that there were most of users in Americas in 2009. On the other hand, Twitter was used all over the world in 2012. Therefore, we reasoned that the reciprocity of Twitter network on 2012 would be smaller value than 2009, because some gaps had occurred between users due to differences of interests, customs or languages.

4.3 Degree of Separation and Diameter

First, we briefly explain the difference of degree of separation and diameter. Both degree of separation and diameter are measures to characterize networks in terms of scale of graph. Degree of separation is given by the average value of the shortest-path length of all pairs of users. On the other hand, diameter is given by maximum value of the shortest-path length of all pairs of users.

The concept of degree of separation was suggested by Milgram who conducted "six degrees of separation" experiment [2][3]. From this experiment, a famous hypothesis widely spread. It was that people could get to know other people all over the world through six or more friends. Lars Backstrom also analyzed degree of separation of Facebook network and presented that the degrees of separation of Facebook is 4.74, which was smaller than they expected.

4.3.1 Experiment Environment

We used HyperANF [6] API as an analysis tool to compute degree of separation and diameter. HyperANF returns an estimation of the number of pairs that can reach within step-t at the t-th iteration, so we can easily compute degree of separation by the iterative calculation. In addition, since the number of iterations until convergence guarantees the lower bound of diameter, we can get the approximate diameter at the same time.

In order to reduce the error, we set the logarithm of the number of registers per counter to 6 (see [6] for details) and ran four times of calculation using HyperANF. Also, it was required a node with large main memory, our computations were performed on a 64-core machine with 512 GB memory that is one of the TSUBAME 2.0 super computer located Tokyo Institute of Technology.

4.3.2 Preprocessing for Experiment

We briefly describe some preprocessing for our experiment in order to obtain the optimized results. We basically used Hadoop and Web Graph APIs [7] in our preprocessing.

1. Prepare user id lists containing serial number from zero to renumber
2. Create edge lists from follower-friend data and renumber with the user id lists.
3. Convert the renumbered edge lists to adjacency lists formatted Ascii Graph [7].
4. Finally, Convert the adjacency lists to compressed data with BV compression API [6][7].

Note that the adjacency lists and the compressed data size are 263GB and 73GB, respectively.

4.3.3 Results and Discussion

Table 3 displays the results of degree of separation and diameter of Twitter network on 2009 and 2012 on each run. First, the average value of degree of separation on 2009 and 2012 are 4.50 and 4.59, respectively. Both values are much smaller than general "six degrees of separation". Comparing these two values, there was only a little difference despite the lapse of three years.

	Degree of separation		Diameter	
	2009	2012	2009	2012
1	4.39	4.48	25	70
2	4.46	4.65	26	71
3	4.53	4.54	25	70
4	4.62	4.71	25	71

Table 3. The results of degree of separation and diameter on each run.

Figure 8 shows the cumulative distribution function of the number of node pairs that can reach within *t* steps. In Twitter network on 2009, 89.2% of node pairs, the path length is 5 or shorter, and for 99.1% it is 6 or shorter. On the other hand, 85.2% it is 5 or shorter, and for 94.6% it is 6 or shorter in Twitter network on 2012.

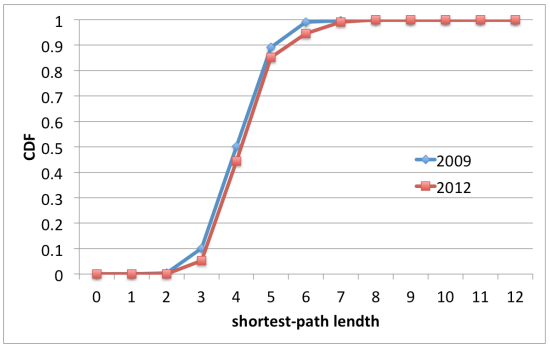

Figure 8. Cumulative distribution function (CDF) of the number of pairs

Also, we got 26 and 71 as the value of diameter of Twitter network on 2009 and 2012, respectively. We assume that one of the reasons that the value of diameter has increased in three years is transition of locality of users as described in Section 3.2. However, its evidence is still not sufficient, there is room for verification. We have to perform a more detailed analysis in the future.

5. CONCLUDING REMARKS

Social network services such as Twitter, Facebook, MySpace, LinkedIn have been remarkably growing over time. Especially, Twitter network structure has changed considerably with the rapid increase of users, so it is significant to analyze the current Twitter network. We collected Twitter user data to analyze by crawling for 3 months from July 2012 to October 2012 and conducted a Twitter network analysis in terms growth by region, scale-free, reciprocity, degree of separation and diameter. Through our experiments, we found that the value of degree of separation is 4.59 in current Twitter network. To contrary our expectation, there was only a little difference between 2009 and 2012 in spite of the rapid growth of network.

Our future work includes a more detailed analysis taking into account the regional or time series in order to clarify cluster property, transition of diameter and so forth.

6. ACKNOWLEDGMENTS

We gratefully appreciate the financial support of JST CREST.

7. REFERENCES

[1] United Nations Statistics Division, http://unstats.un.org/unsd/methods/m49/m49regin.htm

[2] Stanley Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967.

[3] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[4] Haewoon Kwak , Changhyun Lee , Hosung Park , Sue Moon, What is Twitter, a social network or a news media?, *Proceedings of the 19th international conference on World wide web*, April 26-30, 2010, Raleigh, North Carolina, USA

[5] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *ACM Web Science 2012: Conference Proceedings*, pages 45−54. ACM Press, 2012. Best paper award.

[6] Paolo Boldi, Marco Rosa, and Sebastiano Vigna. HyperANF: Approximating the neighbourhood function of very large graphs on a budget. In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, *Proceedings of the 20th international conference on World Wide Web*, pages 625–634. ACM, 2011.

[7] WebGraph, http://webgraph.di.unimi.it/

[8] HubSpot, Stateofthetwittersphere. http://bit.ly/sotwitter , June 2009.

[9] The Apache Software Foundation, Apache Hadoop, http://hadoop.apache.org/