2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

A Study on Twitter User-Follower Network

A network based analysis

VenkataSwamy Martha
@WalmartLabs,
Mountain View, CA, USA
vmartha@walmartlabs.com

Weizhong Zhao, Xiaowei Xu
University of Arkansas at Little Rock,
Little Rock, AR, USA
wxzhao1@ualr.edu, xwxu@ualr.edu

***Abstract*—Substantial percent of global Internet users are now actively use Twitter. In recent times, Twitter has been experiencing explosive growth, attracting celebrities consequently a growing mass of user coverage. Insights of such a social network aid researchers in understanding behavioral dynamics of the society. Though there have been attempts to study social networks, they did not scale to process social networks on the scale of Twitter user-follower network. In this paper we uncovered some of the essential properties of the complete Twitter user-follower network. The properties include degree distribution, connectivity, strength of following relationships, clustering coefficient. Our investigations showed that the Twitter user-follower network follows power-law degree distribution. We found Twitter being a connected network. The strength of the relationships among users is distributed nearly uniform on the scale of 0.0 to 1.0. Nearly 90% of the users possess '0' clustering coefficient, which refers to the least possibility to find communities in the network. In addition to the listed properties, this study found communities of users with high clustering coefficient despite many users with low clustering coefficient. A sample of the communities is validated manually for accuracy. The validation proved that the communities are representing users of similar interests. The communities found from this work yields to friend recommendations, target based advertisements, etc.**

***Keywords*—Twitter, Social network analysis, Behavior analysis, Social media**

I. INTRODUCTION

"Tweeting, do you mean singing?", responded my friends, who do not know Twitter, when I talk about tweets on Twitter. So, what is Twitter?

A. What is Twitter?

Not only in our discussion but also modern world knows Twitter as a social network service, days are gone where only birds tweet. Twitter is a social network website (place) to broadcast your activities, opinions and what are you up to with a minimal effort. Each broadcasting message is called a 'tweet'. There are certain audiences (viewers) for your messages called 'followers'. One follows another's posts and also followed by some other. Thanks to simplicity and powerful mechanism to broadcast tweets in real time, Twitter quickly became dominant social networking tool. Not only people (users) but also news corporations, mass media, auto industries what not almost every business entity started using Twitter to update their consumers with latest offers, products, etc. Therefore, an individual has a page (place) for Twitter to update his/her/its followers using tweets. Because of several technical and usability reasons, a tweet is 140 characters length. The limited length taught users to be creative to express their updates and has been improving oratory skills. Due to small length, a tweet is called as microblog.

We know what Twitter is and how an individual uses it to reach out their audiences. Social computing researchers do not see Twitter as a just social networking service but a source of valuable information of users' activities. A tweet is a user activity and depicts the user's whereabouts and other related information. Social computing researchers are very fascinated to find behavioral patterns from such information. Keeping the information in mind, there has been several research groups working various problems using Twitter. The outcomes from research can trace patterns of disease outbreak, responses to a new product launch, feedback to a new federal policy, etc.

The question raised now is "Can Twitter represents enough number of individuals to support the research outcomes?". Yes, 200 million active users on Twitter (as of Jan 31, 2013) who post around 340 million tweets in a day. Because of the massive number of active individuals, researchers are highly active in exploring insights of Twitter users.

B. Twitter User-Follower Network:

An Individual popularity is measured from the number of individuals follow him/her/it. More the followers means more the popularity. One follows you means you gave permission to him/her to view your posts and you have an option to choose your followers. Therefore, the following relationship represents a social relationship among two individuals. From the relationship, we can connect users among themselves to form a social network. We refer it a Twitter User-Follower network. Though it is a directed network, for simplicity we turn it into undirected network. In the undirected network, there is an edge between two users if one of them follows the other. The network is very useful in understanding users' behavior.

Due to restricting terms of conditions in collecting Twitter data, we make use of Twitter dataset collected in 2009. The data is collected by [1] and made available to public for research purpose. The data is the Twitter user-follower social network constitute of 41.7 million users and 1.47 billion following relationships. There have been several research groups [3] using the dataset to analyze it. As of our knowledge, there is no research group that processed the complete Twitter user-follower network dataset and in this paper we accomplish it by making use of GraphStore [4], a scalable graph storage system. GraphStore is developed on top of Hadoop platform [5] to store big data networks efficiently. In order to process

ASONAM'13, August 25-29, 2013, Niagara, Ontario, CAN

the complete Twitter user-follower network, we had to use distributed system to run network algorithms. Hadoop also served as a distributed system to run the algorithms. We developed required network algorithms in MapReduce paradigm [6] which is necessary to take advantage of Hadoop's robust distributed system.

In this paper we process the Twitter user-follower network to characterize the network properties including degree distribution, connected components, relationship strength distribution, and communities. Due to the size of the network, we run all our experiments on GraphStore. To be focused on network characteristics, we do not mention run time of each experiment.

II. Related Work

This paper is not the first attempt to study Twitter user-follower network. Researchers have been investigating social networks including Twitter to unravel social dynamics. Barabasi and Albert found the long tail degree distribution in social network in 1999 [6]. Such networks are called scale free. Since then many of social network found to be scale free. On the other hand, weak ties found to be strong in social networks [7]. There are several other investigations have been performed on several social networks. Most of such social networks are smaller than scale of current social networks. Given the scale of the social network in recent times, it is hard to manage massive social network representing millions of users and billions of relationships among them. Recent attempt to study a massive social network is published in [2] where the author developed an algorithm to find communities in Twitter user-follower network. In [2], the experiments are performed on a reduced network obtained from eliminating users with more than 900 followers and also users who follow more than 900 users.

Motivated from the lack of ability to process massive networks such as Twitter user-follower network, we performed experiments to study complete Twitter user-follower network through some of the important network metrics in this paper. It is first of its kind to attempt process such a huge network on a small scale computing infrastructure. The network is large enough to consider distributed system for management. Though there is handful of distributed graph processing systems such as Hadoop, there is no efficient graph storage management on Hadoop. We developed a scalable and efficient graph storage system in specific for large graph. Detailed illustration of the storage system is found in [3]. Benefited from the storage system, we processed complete Twitter user-follower network in a given limited computing infrastructure of 30 machines each with 8 core processors and 15GB RAM connected by 10/100MB ethernet cable. This paper focuses on studying the characteristics of the Twitter user-follower network disregarding the implementation details. The implementation details are also provided in [3].

III. Characteristics of Twitter

Relationships among individuals are put together to construct a social network. The edges denote the existence such as. Social networks have been shown to have common characteristics that include positively skewed degree distribution, connectivity and clustering coefficient. Each of the network property of Twitter is computed and discussed in the following.

A. Degree Distribution

The degree of a node in a network is the number of connection it has with other nodes. The degree distribution is then the frequency distribution of different degrees across the nodes in the network. Degree distribution of the Twitter user-follower network is presented in Figure 1.

The plot in the figure suffers from long tail distribution where major users in the network possess low degree. To study the degree distribution further, we plot log-log plot of the degree distribution.

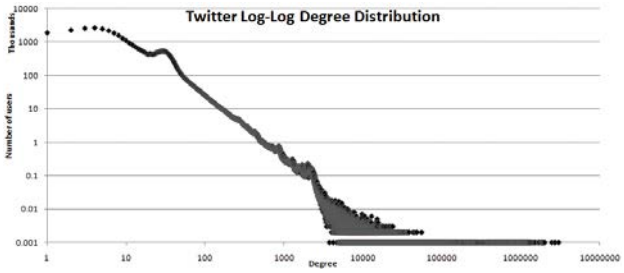

Figure 1: Twitter User-Follower network Log-Log degree distribution

From the log-log plot of the degree distribution, it is clear that the network follows power-law degree distribution as of other social networks.

The maximum degree of the network is 2,997,487. The density of a network is the number of edges divided by number of possible edges and for the Twitter user-follower network, the density is 28.87.

B. Connected Components

A user is "reachable" by another if there exists a set of connections by which we can trace from the source to the target user, regardless of number of hops over the path. Since the Twitter user-follower network is undirected, if a user A reaches another user B implies B also can reach A. If one or some users in the network cannot reach others, the network is constitutes of more than one sub networks. In this experiment we would like to find out the number of sub-networks (connected components) in the Twitter network. Unsurprisingly, the experiment (completed in 14 iterations) outcome showed that the network is a connected component. Therefore, a user in Twitter in can reach other user.

C. Tie strength distribution (SS distribution)

Tie-strength has been in the focus of social science research. The social sciences use the term tie strength to denote this differential closeness with the people in our lives. The strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie [6]. Research from [6] showed the strength of weak ties. It is showed the importance of weak ties in real world especially in labor market. The likely friends of two individuals who are

closely familiar tend to be more overlapping than those of two arbitrarily selected individuals. The familiarity is the strength of the connection between two users. Implied from the statement, the strength of a tie is directly proportional to number of common friends. We can refer the strength as similarity of two users in terms of friends. Therefore,

$$\text{Strength of a tie (A,B)} = \text{similarity of (A,B)} = \frac{|A \cap B|}{\sqrt{|A||B|}}$$

We computed tie strength for each connection in the given Twitter network. It is not feasible to plot the strength values for 1.47 billion connections. So we plotted a strength distribution plot. The strength distribution is the frequency distribution of distinct strengths across the connections in the network. The strength distribution plot is presented in Figure 2.

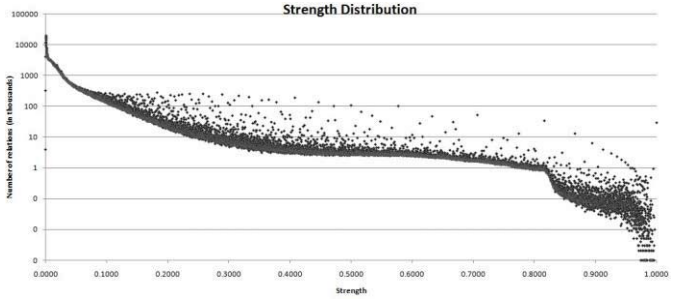

Figure 2: Twitter User-Follower network relationship strength distribution

To have clear picture of the distribution we also plotted cumulative distribution function (CDF) for the tie strength distribution and presented in Figure 3.

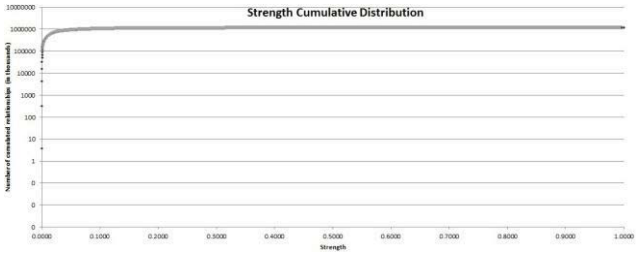

Figure 3: Twitter User-Follower network relationship strength cumulative distribution

D. Clustering Coefficient

Clustering coefficient of a vertex in a network represents how well connected the neighborhood of the vertex is. If the neighborhood is fully connected, the clustering coefficient is 1 and a value close to 0 means that there are hardly any connections in the neighborhood. Formally, Clustering coefficient of a vertex is the ratio of number of connections in the neighborhood of the vertex and the number of connections if the neighborhood was fully connected. In an undirected network, the number of possible connections among vertices in a neighborhood of a vertex 'a' is N(a)*(N(a)-1)/2. Let only y(a) of them be present. The clustering coefficient of this vertex, CC(a) $\equiv$ y(a)/[N(a)*(N(a) − 1)/2], is the fraction of existing connections between nearest neighbors of the vertex. As in earlier cases, here also it is not feasible to plot the clustering coefficients for all the edges. So we plot clustering coefficient distribution in Figure 4.

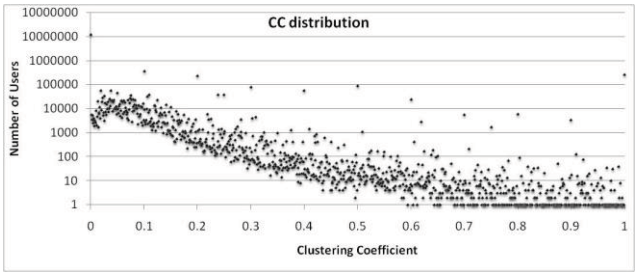

Figure 4: Twitter User-Follower network relationship strength cumulative distribution

Averaging CC(a) over all vertices of a network yields the average clustering coefficient of the network, C. The clustering coefficient is the probability that two nearest neighbors of a vertex are nearest neighbors also of one another. The clustering coefficient of the network reflects the "cliquishness" of the mean closest neighborhood of a network vertex, that is, the extent to which the nearest neighbors of a vertex are the nearest neighbors of each other. From the experiment, we found that Twitter user-follower network's average clustering coefficient is 0.072. From the value, we can say that the network do not possess stringent modular structure.

E. Communities in Twitter

Network clustering (or graph partitioning) is an important task for the discovery of underlying structures in networks. The clusters found in the network represent communities in the social network. In real world people do divide into groups along lines of interest, occupation, age, and so forth, and the phenomenon of assortativity but here with SCAN algorithm we attempt find latent knowledge i.e. community structures from the given network. Many algorithms find clusters by maximizing the number of intra-cluster edges. In this experiment we leverage SCAN (Structural Clustering Algorithm for Networks) [8], which detects clusters in the Twitter user-follower network. The algorithm clusters vertices based on a structural similarity measure. The MapReduce version of the SCAN is PSCAN which is presented in paper [9]. The algorithm finds the structural similarity measure of each edge and filters out the edges that have low measure ($<\varepsilon$). The connected components in the resulting network are clusters from the SCAN. The PSCAN on Twitter found 313,562 communities, at ε=0.6, each of which with different sizes. In an attempt to analyze the communities we plotted a community distribution. The distribution is the frequency distribution of communities across distinct community sizes. The community distribution is presented in Figure 5.

The community distribution plot suffers from long tail property and so we also plot log-log plot of the community distribution and presented in Figure 5.

The log-log plot shows that not only the Twitter user-follower network follows power law but also communities. It again proves the hypothesis from [10].

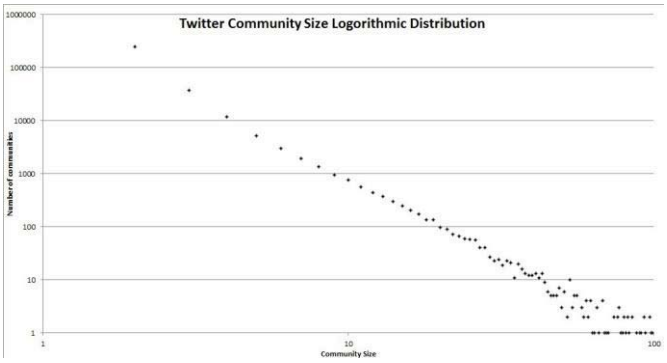

Figure 5: Twitter User-Follower network community size Log distribution

Besides the community size distribution, we also manually validated the communities for their significance. We manually picked 6 communities at random from the 313,562 communities and did lookup Twitter for corresponding users from the communities. Such investigations lead to interesting findings. Each of the community has their significance through representing variety of common interests. For example, a community is representing members from social media based industry while some other community representing members from Berlin, Germany. Through the findings, we can conclude that there are significant group of members in Twitter that form communities. The findings from the investigations are listed in Table 1.

IV. Discussion

It is incontrovertible that Twitter is one of the popular social networks. If Twitter is a nation then it could be a third largest nation in the world. There are 200 million active users in the Twitter. Such a massive social network is certainly a thing of interest for social researchers. Our investigations in the Section II found several interesting facts of the Twitter.

Twitter is highly sparse social network, large portion of the users are not active in following/follower relationships. At the same time there are users with substantially large nearly 2 million followers. Such users are typically celebrities. On average a user follows and followed by at least 28 other users. Though the Twitter user base is in widely from almost all the continents, there is a direct or indirect relationship between any two users. The indirect relationship refers to the user can reach one other user through his neighbor users. As a result of sparse network, we observed most of the relationships are weak. The users are following the users which do not follow or followed the users of similar interest. *Strength of weak ties* law does not stand here as most of the relationships are one way. Because of weak relationships there are hardly communities exist. There are very few communities in such a massive user base. There are nearly 313K communities in the Twitter. A couple of users in a community who do not have follower/following relationship are predicted to be connected in near future. This analogy is applicable for follower/following recommendation. The same approach can be leveraged by industries to incorporate target based advertisements.

On a whole Twitter is a ocean of information and there is a wealth of insights that can improve the way of social research. Our investigations paved a path in the same direction and we plan to extend the investigation to apply it to a case study in a business model.

V. Conclusion

In summary, the Twitter User-Follower network follows power-law degree distribution and has community sizes. We found Twitter to be a connected network, meaning there is a path from each user to any other user. The strength of the relationships among users is distributed nearly uniform on the scale of 0.0 to 1.0. Most of the users possess nearly a 0 clustering coefficient, which refers to the least possibility to find communities in the network. The communities we find in the experiment show communities that follow the power law in size distribution. Most of the users are not part of these communities because of a low clustering coefficient and vice versa. From this we can conclude that the follower-following relationships in Twitter are sparse and result in very few as few as 300,000 communities in the network. The communities found in this investigation are of interest in recommendations, advertising, etc. Motivated by the results, we apply the investigations to a business model in a near future.

Acknowledgment

This project was funded by Acxiom Corporation. This work was supported in part by the National Science Foundation under Grant CRI CNS-0855248, Grant EPS-

Table 1: Listing of randomly picked communities from the observed communities

Community ID (arbitrarily chosen member in the community)	*Time zone*	*Country*	*Common interests*	*Profile Access*
10001792	Santiago	Chile	art, design, graphics	available
10004992	Berlin	Germany	--	protected
10035342	--	--	Social media based managers, architects, marketers, strategist and entrepreneurs	available
10104932	--	--	Adult content	available and most accounts are suspended
10132722	--	--	Teaching and education	available
10139742	--	--	mobile app development	available

0701890, Grant EPS-0918970, Grant MRI CNS-0619069, and OISE-0729792. Weizhong Zhao would like to thank the support of the National Natural Science Foundation of China (No. 61105052). N000141010091). The views expressed in this article are solely those of the authors and do not necessarily reflect the official policies or positions of their employers.

REFERENCES

[1] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, "What is Twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, New York, NY, USA, 2010, pp. 591-600.

[2] Akshay U. Bhat, Scalable Community Detection using Label Propagation & Map-Reduce, Nov 15 2012.

[3] VenkataSwamy Martha, "GraphStore: A Distributed Graph Storage System for Big Data Networks," University of Arkansas at Little Rock, Little Rock, Ph.D. Thesis 2013.

[4] (2012, Nov 15) Apache Hadoop Project. [Online]. http://hadoop.apache.org/

[5] Jeffrey Dean, Sanjay Ghemawat, and Google Inc, "MapReduce: simplified data processing on large clusters," in *In OSDI04: Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation*, 2004.

[6] A. L. Barabasi and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509-512, 1999.

[7] Mark S. Granovetter, "The Strength of Weak Ties," *American Journal of Sociology*, vol. 78, no. 6, pp. pp. 1360-1380, 1973.

[8] Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, and Thomas A. J. Schweiger, "SCAN: a structural clustering algorithm for networks," in *Knowledge Discovery and Data Mining*, 2007, pp. 824-833.

[9] Weizhong Zhao, VenkataSwamy Martha, and Xiaowei Xu, "PSCAN: A Parallel Structural Clustering Algorithm for Big Networks in MapReduce," in *International Conference on Advanced Information Networking and Applications (AINA)*, 2013.

[10] Andrea Lancichinetti and Santo Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Physical Review E*, vol. 80, no. 1, p. 9, 2009.