Link Formation Analysis in Microblogs

Dawei Yin   Liangjie Hong   Xiong Xiong   Brian D. Davison
Department of Computer Science & Engineering, Lehigh University
Bethlehem, PA, USA
{day207, lih307, xix209, davison}@cse.lehigh.edu

ABSTRACT

Unlike a traditional social network service, a microblogging network like Twitter is a hybrid network, combining aspects of both social networks and information networks. Understanding the structure of such hybrid networks and to predict new links are important for many tasks such as friend recommendation, community detection, and network growth models. In this paper, by analyzing data collected over time, we find that 90% of new links are to people just two hops away and dynamics of friend acquisition are also related to users' account age. Finally, we compare two popular sampling methods which are widely used for network analysis and find that ForestFire does not preserve properties required for the link prediction task.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation,Performance

Keywords: link formation, link analysis, microblogs, link prediction

1. INTRODUCTION

The use of online social networks and social media in general has surged in recent years. In this work, we focus on the understanding of link formation of one particular type of social service—that of the microblogging network. In microblog services such as Twitter, Yammer and Google Buzz, participants form an explicit social network by "following" (subscribing to) another user. Unlike common online social networks such as Facebook, LinkedIn or Myspace, a followed user has the option but not the requirement to similarly follow back. Thus, relationships in these social networks may be asymmetric. Thus, User B in a microblog service can generate messages, and any followers of B, such as A, will automatically receive those messages along with messages generated by all other users that A follows. The combination of multiple message intentions and asymmetry of connections has led some to call microblogging services such as Twitter "hybrid networks" [4]. They are hybrid not just because they can carry multiple types of messages, but also because participants create links for multiple reasons—to be social (e.g., to connect online to existing offline social contacts) or to link to an information source [4].

Type	Dynamic	Static	ForestFire
Unknown	0.08126	0.05939	0.90350
⇔⇔	**0.48712**	**0.49710**	0.00028
⇔⇒	0.03974	0.02341	0.00102
⇔⇐	0.04082	0.06068	0.00264
⇒⇔	0.01636	0.05391	0.00155
⇒⇒	0.03706	0.06889	**0.04457**
⇒⇐	0.01112	0.03830	0.01955
⇐⇔	0.17471	0.12875	0.00087
⇐⇒	0.02700	0.00869	0.00456
⇐⇐	0.08477	0.06084	0.02141

Table 1: The distribution of relationship types for new links.

Recently, Golder et al. [2] discuss prediction specifically in Twitter. They analyze several principles for link prediction, such as shared interests, shared followers, and mutuality. Romero and Kleinberg [4] also introduce the hybrid network concept and study the directed closure process in link formation Twitter. In this paper, by analyzing data collected over time, we will uncover more properties for the link formation in ego-centric networks.

2. DATASETS AND ANLYSIS

We randomly sampled 1000 users out of 9,026,165 users active between early February and the end of March 2010. Though users may appear multiple times in the public timeline, we sampled by name, not by tweet, so highly active users had no additional selection advantage. In the end, we had 979 users as our target users.[1] We monitored daily the changes in the selected users' ego-centric networks on Twitter. The number of immediate friends and followers of the 979 target users was nearly 200,000. The data we used in this paper is from April 5th to May 12th, 2010.[2]

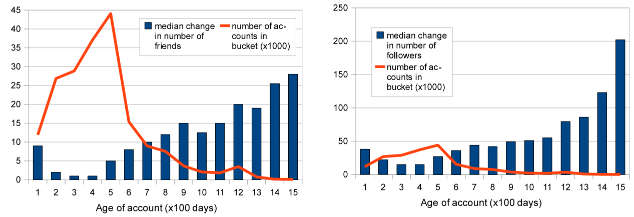
2.1 New Links Analysis

By regularly examining the changing networks, we determine from where new links come. We monitored the changes of ego-networks for each of the 979 users. We collected a total of 18,777 new friends for the 979 users. Most new users are friends of friends. In particular, 17251 (91.78%) new friends were second level neighbors within the target user's ego-network and the remaining 781 ($P_{unknown} = 8.12\%$) new friends were of unknown relationship (i.e., more than two hops away). For each of the new friends, we further check their relationship type with the 979 target users.

SIGIR'11, July 24–28, 2011, Beijing, China.

[1] During monitoring, 21 users changed their privacy setting to "protect", preventing us from continuing to collect their information.

[2] The data is prior to the introduction of Twitter's friend recommendation system which may introduce a link formation bias.



(a) Change in number of friends.

(b) Change in number of followers.

Figure 1: From April to August 2010, the changes in the median number of neighbors as a function of account age.

Table 1 shows the distribution of relationship patterns for each new friend, prior to link creation. For example, ⇔⇔ means the new friends is the reciprocal friend of some reciprocal friend of the target users. Similarly, ⇒⇒ means the new friend is the following-only friend of some following-only friend of the target user. It is similar for other types. Based on the results shown in Table 1, if we use two-hop neighbors in the ego-network as candidates to recommend as friends, we will only miss 8.12% of new friends. Another interesting result is that if the two users share an audience, it does not mean that they are interested in each other. Among two-hop users, most of the new friends have relationships ⇔⇔, ⇐⇔, ⇐⇐. In Romero and Kleinberg [4], they only find ⇒⇒ is an factor of link formation. We extend their observations and find ⇔⇔ is by far the most important indicator for future link formation.

2.2 Age Analysis

We analyze the relationship between the changes in the size of a user's network and user's account age. We compare two snapshots (April 5th 2010 and August 20th 2010) of profiles of 200,000 users. The results are shown in Figure 1. The x-axis is the user's account age and y-axis is the median change in the number of neighbors. We find that in the very beginning (within 100 days), the users add many friends and then for the older users (100-400 days), their friends seem more stable. For much older users (more than 500 days), we find that their number of new friends is larger and larger, not as we expect. For followers, Figure 1(b) shows a *rich get richer* pattern; the older the user, the larger the increase in followers. A more detailed analysis (not shown for space) reveals that young accounts (e.g., less than two years) have a larger (but decreasing over time) change in followers and friends, while more established accounts (from about two years on) have a more consistent relative growth rate.

3. COMPARING SAMPLING METHODS

Link prediction experiments are usually based on sampled graph rather than the whole graph before deploying it on the real system. However, real dynamic data usually is not available, in which case, artificial data is necessary. We study two methods. ForestFire [3] is a popular sampling method, preserving graph properties on the sampled graph, such as some static properties (e.g., degree distributions, clustering coefficients), temporal properties (e.g., shrinking diameters) and cascading properties (shown in [1]). We sample a graph by ForestFire, which contain 1,607,178 users, and remove 10% links at random as test data. Another data set is based on the April 5th snapshot of ego-centric network. For 1000 egos, we remove 10% links at random as test data, which we call Static data. Treating those removed data as the new links, we perform the same analysis experiment as Section 2.1 to analyze whether the artificial data can retain properties consistent with real data. The results are also shown in Table 1. The Static data keeps consistent the distribution among all kinds of the two-hop relationships. For ForestFire, the properties are quite different from Dynamic: fewer than 10% of the candidates are found on the second hop, causing ego-network based structural methods, such as common neighbors and Jaccard's coefficient to fail on such cases.

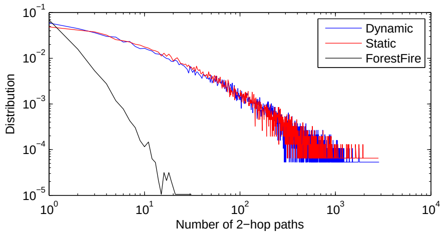


Figure 2: Distribution of links with exactly x 2-hop paths in the graph generated by each sampling method

Furthermore, the distribution of two-hop relationships are also changed. The fraction of ⇔⇔ becomes very low and ⇒⇒ becomes higher. We also draw the distribution of the number of 2-hop paths from the target users to the candidates. Figure 2 shows the distribution of three data sets. We can see that the distribution of Static data is very similar to the Dynamic data, but the distribution of ForestFire is different from real Dynamic data. That means that even when we only use the candidates who are still available via two hops from the target users as our test data, the algorithms may still generate different performance, compared with the real Dynamic data. These experimental results suggests that for the link prediction task, the common evaluation method which is based on ForestFire sampled data may not cause the same the results as real data. Finally, we run a simple but popular method—Jaccard's coefficient—on the three data sets.[3] The results are 0.116 in Dynamic, 0.071 in Static, and 0.0013 in ForestFire data respectively. As we expect, it totally fails in the ForestFire data set and the performances on Dynamic and Static are similar.

4. CONCLUSION

By analyzing data collected over time, we find that 90% of new links are to people just two hops away and the dynamics of new link creation are affected by the user's account age. Based on these properties, we compare two methods of collecting network data—ForestFire and Ego-network sampling. The results show that ForestFire does not preserve important properties of the ego-network and is thus not suitable for the link prediction task.

Acknowledgments

This work was supported in part by a grant from the National Science Foundation under award IIS-0545875 and an equipment grant from Sun Microsystems.

5. REFERENCES

[1] M. De Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher. How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media? In *ICWSM 2010*.

[2] S. Golder and S. Yardi. Structural predictors of tie formation in Twitter: Transitivity and mutuality. In *SocialCom 2010*.

[3] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD 2006*, pages 631–636.

[4] D. Romero and J. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *ICWSM 2010*.

[3] We use the standard F1-measure in the break even point.