From Strangers to Neighbors: Link Prediction in Microblogs using Social Distance Game

Dawei Liu[§], Yuanzhuo Wang[†], YanTao Jia[†], Jingyuan Li[†], Zhihua Yu[§†]
[§]Institute of Network Technology, Institute of Computing Technology (Yantai), CAS, Beijing, P. R. China
[†]Institute of Computing Technology, CAS, Beijing, P. R. China
liudw@int-yt.com, {wangyuanzhuo, jiayantao, lijingyuan, yzh}@ict.ac.cn

ABSTRACT
Link prediction is an important research topic for social network. In this paper, we propose the notion of social distance considering both structural and interactive characteristics, to measure the closeness among a group of people in Microblogs. We then model the procedure of link prediction with a coalitional game in a directed graph under the concept of homophily. We explain the solution concept of generating predictive future neighbors for a given agent and propose a weighted social welfare maximization solution for social distance formation. Experiments were applied over a Twitter dataset of 140,000 users and 400,000,000 tweets, and the results testified the effectiveness of our game theoretic approach in predicting the likelihood of future associations between people.

Categories and Subject Descriptors
J.4 [**Computer Applications**]: Social & Behavioral Sciences

General Terms
Algorithms, Experimentation

Keywords
Link prediction, Social Distance, Microblogs, Coalitional Game

1. INTRODUCTION
Social networks have been studied extensively in the context of analyzing interactions between people and exploring the structural properties in those interactions. Link prediction is an important task which leverages either the structure of the network or the attribute information at different agents to determine or predict future links. Link prediction also has many applications in different domains like information retrieval, bioinformatics and e-commerce. There exist a variety of structural and relational models in the literature for link prediction, ranging from feature-based classification and kernel-based method to matrix factorization and probabilistic graphical models [1].

We consider link prediction in online social networks, specifically in the popular Microblogging platforms such as Twitter and Sina Weibo, where the associations between agents are represented in the notion of *follow*. We focus on the problem of *from strangers to neighbors*, that is, given an agent, the likelihood of its future neighbors is to be predicted. Note that the concept of neighbors here stands for the set of agents it follows. Link prediction has been studied far before the bloom of Internet in the field of sociometry. Leskovec et al. [2] analyzed an activity network based on user interaction in a large instant messaging network. They discovered that the activity network displayed strong influence of *homophily* in interaction, where similar users interact with each other in a considerably higher frequency, which means that similarity breeds connection and people tend to form communities with their own kind. The concept of *homophily* reveals the small world phenomenon, the principle that most of us are linked by short chains of acquaintances. Kleinberg's study [3] shows that individuals operating with purely *local information* are very adept at finding these short chains.

Most of the existing models formulate link prediction by a learning problem with a graphic representation, where a vertex represents an agent and an edge between two vertices represents the interaction between the two agents. The learning paradigm in this setup typically extracts the similarity between a pair of vertices by various graph-based similarity metrics and uses the ranking of the similarity scores to predict the link between two vertices [1]. Hasan et al. [4] considered the link prediction in a supervised machine learning setup, where a binary classification task is performed. They showed that using external data outside the scope of graph topology can significantly improve the prediction result, and then provided a comparison of various similarity metrics as features, such as the sum of neighbors, the shortest distance, the keyword match count etc. Different structural and relational models [5] [6] have also been proposed in the literature for link prediction. While prominent in modeling various social problems, game theoretic concepts have been ignored in link prediction. Recent years, the concepts of cooperative game theory, namely coalitional games attract more attentions in the academic world. Coalitional game shows how a group of self-interested players interact with each other to get more payoffs than they could achieve individually. Scott Shenker once said that "The Internet is an equilibrium, we just have to identify the game" [7]. We therefore believe that with an appropriate model, the problem of link prediction in social

Conference'10, Month 1–2, 2010, City, State, Country.

network can also be solved in a game theoretic way. Few game theoretic methods have been proposed to solve the link prediction task. Fabrikant et al. [8] introduced a local connection game in which the edges have constant cost and the agents try to minimize their cost plus the sum of distances to all other pairs. However, the utility function partly considers a global objective which minimizes the distance to all pairs and may not be practical in Microblogs [1].

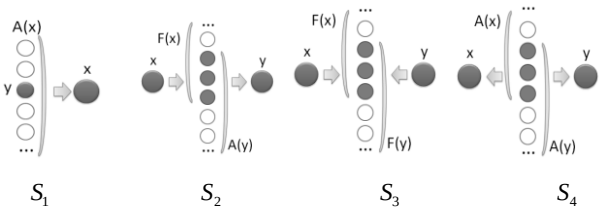
We hold the idea that social networks exhibit *homophily* and that the agents prefer to create ties with other agents who are *close* to them. We analyze the interaction patterns of agents in Microblogs and extract a set of local agents as candidates. The local structure is then formulated with a directed weighted graph, and the notion of social distance is introduced to represent the *closeness* among agents. Finally, a sequence of certain agents, or potential neighbors, for the given agent is generated by a coalitional game framework.

2. SOCIAL DISTANCE GAME

2.1 Motivation

To measure the closeness among people of a group, we present the definition of social distance which represents both structural and interaction characteristics. Firstly, social distance exists between every agent pair in the group. Some agent pairs, although without existing relations for now, may have latent associations in the future and should be studied by the link prediction models. The formation in social network is more than building a graph simply based on linkage information. **How to quantify the distance between disconnected agents** is a key issue. Secondly, in real world, social metric violates the triangle law, which can be taken as another view of the impact of dynamic of social networks. With the evolution of networks, some new links formed from latent associations. The change of the network structure leads to the mismatch of the original distance between agent pairs. As a result, any solutions based on former static states become meaningless. **How to design the social distance in a persistent and comparable form that can be robust to the dynamic evolutions** is another important problem.

We solve the two aforementioned aspects in our social distance game model by using a novel social distance definition. The information captured in microblogs is divided into two categories: structural features and interaction features. Structural features consist of relations between agents such as follow, neighbors and so on, which can be represented by a certain graph. Interaction features are content communications between agents, such as @, comments, interests, topics, etc., which can also be used to detect inherent associations.



S_1 S_2 S_3 S_4

S_1 **:Agents follow x**
S_2 **:Agents followed by the neighbors of x**
S_3 **:Agents with the same neighbors of x**
S_4 **:Agents with the same followers of x**

Figure 1. Interaction patterns between an agent x and a candidate friend y in local structure

2.2 Social Distance

Linkage data in microblogs are represented with a weighted and directed graph $I=(N,E)$, where $N=\{x_1,...,x_n\}$ is the set of agents, $E=\{e_{ij}\}$ is the set of edges. Accordingly, E is generated by the relations of *follow*: each element e_{ij} is a directed link from agent x_i to agent x_j, indicating that agent x_i follows agent x_j in microblogs. For an agent x_i, the set of agents it follows (neighbors) is $F(x_i)=\{x_j\in N \,|\, e_{ji}=1\}$; the set of agents who follow it is $A(x_i)=\{x_j\in N \,|\, e_{ji}=1\}$. As shown in figure 1, interaction between an agent *x* and a candidate friend *y* has four types. Agents that follow *x* may communicate with *x* actively; Agents that followed by neighbors of *x* are those "friends of friends" and may be introduced to *x* through intermediate agents. Agents with same neighbors or followers of *x* may share the same interests or involved in the same community.

Therefore we can get a local structure $I'=(N',E')$ (sub-graph) for a given agent *x* (figure 2) according to the interaction patterns, in which the candidate agents set of future neighbors is defined as follow:

$$N'=\{x\}\cup A(x)\cup F(A(x))\cup F(x)\cup F(F(x))\cup A(F(x))$$

Then we represent the interaction sub-graph $I'=(N',E')$ with a weight matrix $W=[w_{ij}]$, where $w_{ij}\geq 0$ for any edge e_{ij}, and that $\sum_j w_{ij}=1$ for any i. The weight w_{ij} is determined considering the interaction between agent x_i and x_j. In the time interval $(t_0,t_1]$, the amount of communication from agent x_i to x_j is denoted by $CO(x_i,x_j)$. Therefore the total amount of communication from agent x_i is $CO(x_i)=\sum_{x_j\in N, i\neq j} CO(x_i,x_j)$, which corresponds to the whole agent set N containing agents not in the sub-graph agent set N'. Hence, we assign the self-loop weight w_{ii} to represent the total amount of communication from agent x_i to agents not in the reduced candidate agents set of future neighbors:

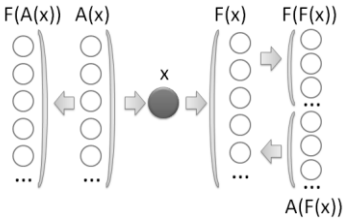


Figure 2. Local structure of an agent x

$$w_{ij} = \frac{1}{CO(x_i)} \begin{cases} \sum_{x_k \notin N', i \neq k} CO(x_i, x_k) & if \quad i = j \\ CO(x_i, x_j) & if \quad i \neq j \end{cases}$$

In the induced agent set N' of a given agent x, the social distance to a stranger $\{y \notin A(x) \bigcup F(x)\}$ is determined by the *structural* and *interaction* features of all his neighbors and followers $\{z \in A(x) \bigcup F(x)\}$. Suppose there are totally m neighbors or followers, and totally n strangers in the set N', we sort the matrix $W = [w_{ij}]$ with index order: $[x, f_1, ... f_m, s_1, ... s_n]$. Thus we have the definition of **social distance** in a triple form:

$$SD(x, f_i, s_j) = (\sum_{k=x,\, fi...\, fm, s1...sn, k \neq fi} w_{k,fi}) \cdot \min\{path(f_i \to s_j) \quad in \quad I\}$$

In which the sum of $w_{k,fi}$ captures the *interaction* characteristics and is combined with the minimum hops (range from 1 to 3) from agent s_j to agent f_i in the graph $I = (N, E)$ without considering the direction, which represents the *structural* feature. The notion of social distance indicates the closeness between the given agent x and a stranger agent s_j from the aspect of intermediate agent f_i in the local structure $I' = (N', E')$.

2.3 Coalitional Game Framework

A coalitional game model is considered to be a solution process with the definition of social distance. For a given agent x, we construct a local structure containing three types of agents: the agent x itself, the agent set of x 's followers and neighbors and the set of stranger agents. The aim of the game theoretic model is to find a proper strategy to make the coalition get best performance under certain solution concept. In our scenario, the utility for each agent in a coalition is designed to represent *homophily* according to the given agent x, that is, utility in our model is considered from more subjective viewpoints.

We define the **social distance game** $G(I)$ to be the pair $\langle N, v \rangle$ where N is the set of agents defined by I and for any $N' \subseteq N$, $v(N') \subset R^{|S|}$, such that for agent x, $v_x(N') = u(x, N')$. The **utility** of the candidate agent s_j for the given agent x in coalition $N' \subseteq N$ is:

$$u(s_j, N') = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{SD(x, f_i, s_j)}$$

The **social welfare** is defined as the sum of utilities of selected $k (k \leq n)$ agents:

$$SW(x) = \sum_{k \leq n} u(s_j, N')$$

We solve the social distance game in the concept of social welfare maximization, that is:

$$\{s_j\}_k^* = \arg\max\{SW(x) = \sum_{k \leq n} u(s_j, N')\}$$

The k agents which make the maximum social welfare for the given agent x are considered to be the results for link prediction task in our scenario.

3. EXPERIMENTS

The Twitter linkage dataset were crawled through the API service provided by the official Twitter website. By randomly selecting 10,000 Twitter users, updating their immediate neighbors per day from the period of Oct. 1st and Nov. 19th, 2012, we built the *structural sub-graph.* Meanwhile, we extracted the interaction data -- tweets of these users per day -- and used them to construct the *interaction sub-graph*, where agent x have relations with agent y if x 's tweet contains the syntax @ y or RT @ y, or equivalently, x retweets y or mentioned y in its tweets. In total, there are 140,000 users and 400,000,000 tweets. We randomly selected 1,000 pairs of snapshots of the dataset, and used the first snapshot to predict the following links in the second snapshot. The interval between these two snapshots is one week.

We selected four representative quantities which have been proven [5] to perform reasonably well in previous studies to compare with our proposed approach.

- Common neighbors.

The number of followers and neighbors that agent x and y have in common: $CN(x, y) \equiv |\Gamma(x) \bigcap \Gamma(y)|$, where $\Gamma(x) = A(x) \bigcup F(x)$.

- Jaccard's coefficient.

Defined as the size of the intersection of the neighbors of two nodes, $\Gamma(x)$ and $\Gamma(y)$, divided by the size of their union, characterizing the similarity between their sets of neighbors:

$$J(x, y) = |\Gamma(x) \bigcap \Gamma(y)| / |\Gamma(x) \bigcup \Gamma(y)|$$

- Adamic-Adar.

A refinement of $CN(x, y)$ by weighting common neighbors based on their degrees, instead of simple counting. Therefore the contribution is penalized by the inverse logarithm of their degree:

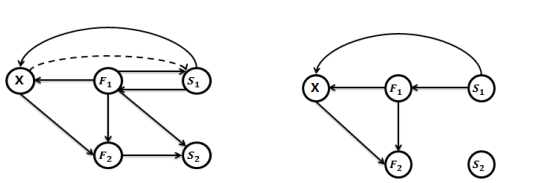
$$AA(x, y) = \sum_{z \in \Gamma(x) \bigcap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$

- *Katz*.

Summation over all possible paths from x to y with exponential damping by length to weight short paths more heavily:

$K(x, y) = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{x,y}^l|$, where $paths_{x,y}^l$ is the set of all paths with length l from x to y.

We first diagrammatize the effectiveness of our social distance game to capture the *interaction information* in Microblogs through a subset of agents from the real world data as an example shown in figure 3. For the given agent x, we construct the structural sub-graph and the interaction sub-graph. In the example, the index order is $[x, f_1, ... f_m, s_1, ... s_n] = [x, f_1, f_2, s_1, s_2]$. We choose the value of k to be 1 in our social distance game, which means we select one agent from s_2. We calculate



structural sub-graph **interaction sub-graph**

Figure 3: An example of sub-graphs

Table 1. Experimental results: Similarity

similarity	*Common Neighbors*	*Jaccard*	*Adamic-Adar*	*Katz*	*Our's*
(X,S1)	1	0.5	1.661	0	**1.67**
(X,S2)	**2**	**1**	**3.757**	**0.0025**	1.627

similarities between agent x and agents $\{s_1, s_2\}$ by using the above four similarity measures and our measure and list these values in Table 1.

As is shown in Table 1, our strategy to calculate the similarity between x and s_1 obtains the largest value in the first row. This means for the agent s_1, our measurement performs the best, which is consistent with the ground truth since x interacts more often with s_1 than with s_2. This reveals that our method indeed characterize the truth that the formation of future link in Microblogs not only relies on the structural features of the network but also relies on the interaction behaviors between agents. Specifically, although agent s_2 has more topological features in common with the given agent x, it makes no communications in the local structure in the snapshot. Our method captured the interaction features with the definition of social distance in the form of using weights of intermediate agents.

Then we calculate the F-measure for each method to evaluate the performance of both precision and recall. As shown in Table 2, our method outperforms all the other measures Structural measures such as common neighbors, Jaccard's coefficient and Adamic-Adar get F-measures for the reason that no interaction information is captured. Katz performs worst in our experiment because the vertices on the path considered by the Katz has low degree on the local structure.

Furthermore, we observed that the size ratio of interaction sub-graph and structural sub-graph had influence on the precision of our proposed method. The reasons for this phenomenon are twofold. Firstly, if too few communications between pairs have been captured during the time interval corresponding to the local structure, the impact of $\sum CO(x_i)$ will be trivial. Secondly, a large amount of captured communications were with agents outside the local structure, which means the self-loop weight $\sum_{x_k \notin N', i \neq k} CO(x_i, x_k)$ was too large, which was ignored when calculating the social distance. Therefore, increasing the size ratio of interaction sub-graph and structural sub-graph can help the social distance effectively capture the individual impact of each intermediate agent and improve the prediction precision accordingly.

4. CONCLUSION AND FUTURE WORK

In this paper we focused on the problem of link prediction in Microblogs, and proposed the notion of social distance based on the interaction patterns. Then we proposed a social distance game model with a solution of weighted social welfare maximization. The main idea is that social networks exhibit *homophily* and that the agents prefer to create ties with other agents who are *close* to them. The experimental results on a Twitter dataset proved the efficiency of our game model of predicting the likelihood of future associations between people in Microblogs.

In the future we will continue to use the notion of social distance as a measure for representing closeness between agents in social networks, and discuss different solution concepts of the proposed social distance game corresponding to different scenarios in social network.

Table 2. Experimental results: F-measure

Method	*F-measure*
Common Neighbors	0.0782
Jaccard's coefficient	0.0816
Adamic-Adar	0.1002
Katz	0.0537
Social Distance Game	**0.1529**

5. ACKNOWLEDGMENTS

This work is supported by National Grand Fundamental Research 973 Program of China (No. 2013CB329602), National Natural Science Foundation of China (No. 61173008, 61232010, 60933005), National Science supported planning (No. 2014AA012304), and Beijing nova program (No. Z121101002512063).

6. REFERENCES

[1] M. Al Hasan and M.J. Zaki. 2011. A Survey of Link Prediction in Social Networks. In *Social Network Data Analytics* , Elsevier, 243-276.

[2] J. Leskovec and E. Horvitz. 2008. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web (WWW)*, 915-924.

[3] J.M. Kleingerg. 2000. Navigation in a small world. *Nature*, vol. 406, 24(Aug. 2000), 845.

[4] M.A. Hasan, C. Vineet, S. Salem and M. Zaki. 2006. Link prediction using supervised learning. In *Proceedings of SDM Workshop of Link Analysis, Counterterrorism and Security*.

[5] D. Liben-Nowell and J. Kleinberg. 2007. The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology*. Vol. 58,(May. 2007), 1019-1031.

[6] B. Taskar, M.F. Wong, P. Abbeel and D. Koller. 2003. Link prediction in relational data. In *Proceedings of Neural Information Processing Systems*. Cambridge, MA: MIT Press. 659-666.

[7] N. Nisan, editors. 2007. *Algorithmic Game Theory*. Cambridge University Press.

[8] A. Fabrikant, A. Luthra, E. Maneva, C.H. Papadimitriou and S. Shenker. 2003. On a Network Creation Game. In *Proceedings of the twenty-second annual symposium on principles of distributed computing*, 347-351.