# LSDH: A Hashing Approach for Large-Scale Link Prediction in Microblogs

**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, jiayuantao, lijingyuan, yzh}@ict.ac.cn

### Abstract

One challenge of link prediction in online social networks is the large scale of many such networks. The measures used by existing work lack a computational consideration in the large scale setting. We propose the notion of social distance in a multi-dimensional form to measure the closeness among a group of people in Microblogs. We proposed a fast hashing approach called Locality-sensitive Social Distance Hashing (LSDH), which works in an unsupervised setup and performs approximate near neighbor search without high-dimensional distance computation. Experiments were applied over a Twitter dataset and the preliminary results testified the effectiveness of LSDH in predicting the likelihood of future associations between people.

## 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 (Liben-Nowell and Kleinberg 2007) 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. 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 (Hasan et al. 2011). Most of existing works focus on the formation of association pattern between two agents and model the structural information in a supervised or unsupervised way. Few works considered the large scale problem of real online social network. Compared with other network dataset such as co-author or biological networks, social network especially Microblogs are much more complicated. Different types of relationships must be considered and treated differently.

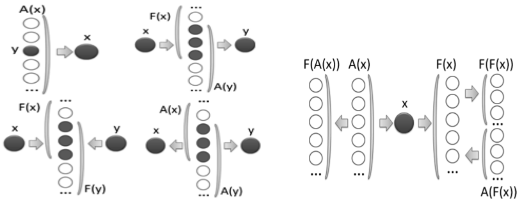
## Methods

We first introduce a novel notion of social distance to represent the closeness among agents, which captures the local structure with interaction information and is different from existing work (Zhou et al. 2009). Then, we propose a fast hashing approach called LSDH, which works in an unsupervised setup and performs approximate near-neighbor search without high-dimensional distance computation in social distance space.

### Social Distance

Linkage data in Microblogs are represented with a 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 $F(x_i) = \{x_j \in N | e_{\overline{ji}} = 1\}$; the set of agents who follow it is $A(x_i) = \{x_j \in N | e_{\overline{ji}} = 1\}$. There are several types of interaction between an agent $x$ and a candidate friend $y$, as shown in Figure1(a): 1) Agents that follow $x$ may communicate with $x$ actively; 2) Agents that followed by neighbors of $x$ are those "friends of friends" and may be introduced to $x$ through intermediate agents; 3) 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 for a given agent $x$ according to the interaction patterns, in which the candidate agents set of future neighbors is defined as follow: $N' = \{x\} \cup A(x) \cup F(A(x)) \cup F(x) \cup F(F(x)) \cup A(F(x))$



(a) Interaction Patterns (b)Local Structure

*Figure 1. Social Distance Components*

(Figure1(b)). In the induced agent set $N'$ of a given agent $x$, the social distance to a stranger $\{y \notin F(x)\}$ is determined by the *structural* and *interaction* features of some or all his local structure. 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 $i$ .Each agent in the target agent's local structure performs as a candidate coordinate $c_i$ of the social distance corresponding to the target agent. The definition for one dimension of social distance in triple form as follows $SD(x, c_i, y) = (\sum_{k \in N'} w_{k,c_i}) \cdot \mathit{shortestpath}(x, y)$ .In which the sum of $w_{k,c_i}$ captures the *interaction* characteristics and is combined with the shortest path (minimum hops) from agent $y$ to agent $x$ in the directed graph $I = (N, E)$, which represents the *structural* feature. The notion of social distance indicates the closeness between the given agent $x$ and a stranger agent $y$ from the aspect of intermediate agent $c_i$ in the local structure. Suppose we choose $C$ agents in $x$ 's local structure to be coordinates of social distance. The overall representation of **social distance** is as follow: $SD_{x,y} = [SD(x, c_1, y), ..., SD(x, c_C, y)]$ where $c_1, c_2, ...c_C$ are agents belonging to $x$ 's local structure.

## Locality-sensitive Social Distance Hashing (LSDH)

When performing link prediction using social distance for a given agent, we first generate a local structure according to the agent's structural linkage information as candidate coordinates. And then we choose all or a subset of candidate agents to calculate interaction weight matrix and determine each dimension. Then we leverage locality-sensitive hashing (LSH) scheme $h_{a,b}(p) = \left\lfloor \frac{a \cdot p + b}{W} \right\rfloor$ with parameterizations. After learning parameters with some samples, an representing $L$ with $k$ : $L(k) = \frac{\log \delta}{\log(1 - p(1)^k)}$, we solve the three parameters $W, k, L$ with the minimum number for average size of each buckets (Datar et al. 2004). We retrieve the given agent in the form of social distance under Euclidean norm. All stranger agents in multiple buckets which the given agent hashed into are retrieved to be the candidate future friend for the given agent.

Table 1. New Links Statistics and Experimental Results

| ***Candidate Type*** | *F(x)* | A(x) | F(A(x)) | F(F(x)) | A(F(x)) | **LS** |
|---|---|---|---|---|---|---|
| *Ave. Num* | 72 | 63 | 3263 | 3874 | 4282 | **8487** |
| *Proportion%* | / | 9.6 | 34.5 | 58.2 | 41.8 | **90.8** |
| *Interaction%* | 6.5 | 9.7 | 11.9 | 48.3 | 15.3 | **91.7** |
| ***Precision%*** | **16.5** | **12.3** | **20.7** | **23.4** | **19.7** | **26.7** |
| ***Recall%*** | **3.5** | **8.2** | **5.8** | **5.3** | **5.8** | **10.7** |

## Experimental Results and Conclusion

We crawled the Twitter linkage dataset through the API service provided by the official Twitter website. In total, there are 12,000 users and 20,0000 tweets. We did a statistical analysis on the new links as shown in Table 1 and found that about 9.2% new links are from those agents "outside" the local structure (LS). What's more, a majority of new links are within two hops from the given agent, where interactions play an essential role in facilitating the growth of neighbor number. We then compare the performance of different coordinates of social distance. We respectively take the whole local structure and several subset of it to calculate the precision and recall of link prediction task, and also the interaction coverage. The results show that with the increase of interaction coverage, there is a significant increase of the precision. It can be seen that the precision of link prediction is improved by taken into account the interaction information. While the recall is related with other agents which are not in the local structure but have a certain amount of interaction. This is also in accordance with the new link statistics. Interaction produced anywhere in the social network, finding future links within any local structure will suffer a loss on recall.

As a conclusion the core contribution is a new general framework combines a multi-dimensional social distance representation and a hashing approach for fast similarity search in the generated multi-dimensional space. The proposed LSDH uses static local structure as candidate coordinates and measures global agents in a high dimensional vector space through dynamic interaction information, which in some sense captured the communication patterns of agents in the social network during a period of time.

## Acknowledgements

This work is supported by 973 Program (2014CB340401, 2013CB329602), 863 Program (2014AA015204), NSFC (61173008, 61232010, 61303244, 61370132), Beijing nova program (Z121101002512063).

## References

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.

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

T. Zhou, L. Lv, Y.C. Zhang, 2009. Predicting missing links vial local information. *Eur. Phys. J.B.* 71(623).

M. Datar, N. Immorlica, P. Indyk and V. Mirrokni. 2004. Locality-sensitive hashing scheme based on p-stable distributions," in *Proc. ACM Symposium on Computational.*