User Preference Based Link Inference for Social Network

Yuqing Sun*†, Haoran Xu*, Elisa Bertino‡ and Demin Li*
*School of Computer Science and Technology, Shandong University, Jinan, China
†Engineering Research Center of Digital Media Technology, Ministry of Education of PRC
‡Department of Computer Science, Purdue University, West Lafayette, USA
Email: sun_yuqing@sdu.edu.cn, hr_xu1990@163.com, bertino@purdue.edu, lidemin1014@126.com

***Abstract*—In this paper, we focus on the link predication problem in social networks. Our approach is based on the observation that there is a large amount of social behavior taking place every day which contains substantial information about user intrinsic characteristics that influence the dynamics of social networks. In order to obtain a deeper understanding of user behavior, we introduce the concept of latent factor to capture the motivation behind social activities. Since user relationships are often asymmetric, we also take into account bilateral user wishes with respect to friend as preferences, which is beyond traditional approaches or overall measurements. Two combination modes are proposed, independent fusion and interdependent fusion, to integrate these hybrid metrics with traditional measurements for link inference. In order to quantify the sensitivity of each element in metrics we use information theory. Experimental results on several real datasets show that our approach has better performance than previous methods.**

I. Introduction

Web based social network services are today very popular. Social networks are usually represented as graphs, in which a node represents a person associated with some attributes and a link represents a relationship between users. Link inference is a critical technique in social network analysis in order to find missing links or predict links that will appear in future, which is often provided to users as a friend recommendation services.

Several approaches have been proposed for link inference based on the analysis of the network structure. Such approaches compute a score as the closeness of nodes based on common friends between two nodes [9] or the paths connecting them [1]. However, such approaches do not take into account node attributes that do influence user social activities. For example, there are 1.2 billion registered users on Facebook who submit personal information such as name, age, sex, occupation, education, hobby, etc. To take node attributes into account, some models use node attributes, such as location [23], social circle [29] or interest group [11], as a reference for link inference [5]. Since this information is static, such approaches have limitations in reflecting node and link changes in social activities. However, the dynamics is the most important characteristic of social services. For example, according to statistics by Facebook[1], over 802 million people log in the system daily, over 300 million photos are uploaded, 510 comments are posted, and 293,000 statuses are updated every minute. A variety of research efforts have thus been devoted to analyze these social behaviors [25] such as common topics or public opinion detection. However, such approaches focus on group activities and do not consider user personalized requirements on social services. Since users often have preferences on making friends and social activities, these methods are not appropriate for link inference.

Another shortcoming of behavior analysis techniques is that only explicit social contents are considered, such as posts, comments, links of web pages etc., while the intrinsic factors behind social activities are rarely investigated. Although such explicit information is very useful, it is not considered to be always reliable as network topology. However, user social behaviors contain much intrinsic information on people social purposes or emotion, which are stable and reliable. The goal of our work is thus to investigate intrinsic characteristics and user social preferences that influence links. Our novel contributions can be summarized as follows.

First, we introduce the concept of latent factor to analyze user behavior. According to psychology and social science, social behaviors are dominated by some latent factors such as cognition, emotion, interest, desire, demand, ideals, beliefs and social values. [2] We then extract the intrinsic factors behind social activities as a metric to assign a score to a potential link. Second, we show how user personalization can be taken into account in link inference. On one side, an entropy based metric is introduced to quantify user social bias on friend attributes and behaviors. On the other side, user bilateral wishes are taken into account when evaluating a link, which is a novel metric with respect to conventional link inference approaches. Third, we propose two combination methods, referred to as independent fusion and interdependent fusion, to semantically combine these hybrid metrics with the traditional network structure based metrics for link inference.We apply information theory techniques to quantify the sensitivity of each considered metric. A potential link is then quantitatively formalized as the closeness between users, which is measured against our metrics. Finally, experiments are performed on several real data sets. The results show that our metrics

[1] https://zephoria.com/social-media/top-15-valuable-facebook-statistics

[2] http://wiki.mbalib.com/wiki/Behavior

978-1-5090-2675-3/16 $31.00 © 2016 IEEE
DOI 10.1109/ICWS.2016.32

outperform previous methods.

The rest of the paper is organized as follows. Next section discusses related work. Then, we formally define the link inference problem and present an overview of our approach. In sections IV and V, we discuss user preference on friend attribute and social behavior, as well as related metrics, respectively. In section VI we propose two strategies for combining multiple metrics. Section VII presents experimental results on real datasets. Finally, we conclude the paper.

II. Related work

Link inference, also called link predication, has attracted increasing attention in recent years. The simplest framework for most link inference approaches is based on the similarity between two users [28]. The more similar the two users are, the more likely is that there exists a link between them. These similarity-based link inference methods can be further classified as structure-based and node-based.

Structure-based Link Inference Methods. In social networks, the structural features of networks are often public and easy to obtain, which explain why many methods mainly focus on structural similarity. The most popular method is Common Neighbors (CN) [17]. The approach by Newman is based on the fact that a positive correlation exists between the number of common neighbors and the probability that they will collaborate in the future [20]. Since the neighbors can reflect user preference and affect user behavior, several approaches have extended the metric by Newman by taking into account various factors, such as CN normalization, topological overlap, 2-distance friends [2], [18].

Besides neighbor information, the paths connecting two users are also used for link inference. The number and length of such paths are both important factors. The more paths exist and the shorter each path, the higher the probability of a link. Katz et al. compute all paths between two nodes and restrict the influence of a long path by an exponential factor [10]. Papadimitriou et al. compute a possible link based on counts of varying length paths connecting two users [21]. Other approaches adopt different kinds of methods for calculating the length of paths, such as random walk [13], [8], co-occurrence probability [27], Markov random network [4].

Many methods recent treat the link inference problem as a clustering problem, according to which users classified in the same group have higher probability of connecting [15]. Machine learning methods are used for clustering users into different groups [34], such as logistical regression [12], support vector machine (SVM) [16] and deep belief network [14].

As these methods only analyze the network structure for link inference, they are unable to leverage the wealth of information about node attributes and dynamic social activities, which are inherent characteristics influencing user links.

Node-based Link Inference Methods. As individuals tend to communicate with other individuals that have similar characteristics, such as educations or interests, several methods take such characteristics into account in order to compute similarities between users. Some methods utilize user interest to evaluate user similarity [29], [11]; location [23] and the set of keywords[3] are also considered important attributes to infer social ties. Some methods thus combine attribute based evaluation with network structure analysis [7], [30]. Gong et al. propose an attribute-augmented social network model which considers each attribute as a node and the user who holds this attribute is added via a link to this attribute node. Based on such extended social network structure, such method has higher accuracy [5]. However, these methods consider neither the dynamics of social behaviors nor user preferences, which are actually important characteristics of social networks and thus critical for improving the performance of link predication.

Besides, user behaviors can also indirectly reflect user preferences. Sun et al. propose a topic modeling framework for social networks, which considers both text and structure information [25]. Qiu et al. calculate temporal features from the time series to characterize behavior evolution, which is used to improve the link prediction accuracy [22]. But such methods focus on topic detection or public opinion mining, in which common interests are the main target and user personalized requirements are not taken into account. Also they only utilize explicit social contents, such as posts, comments, and links to blog and web pages, without considering intrinsic factors behind social activities, which are the real factors affecting user social behaviors.

III. Problem and Framework

A social network is represented as an undirected graph $G\langle V, E, A, B\rangle$, where V is the user set, $E = \{(u, v)|u, v \in V\}$ represents the set of user relationships, A and B represent user attributes and social behavior, respectively. Each user $u \in V$ is associated with a set of attributes; such set is formalized as a vector $A_u = \{a_1, a_2, ..., a_n\}$, where n represents the number of total attribute values. $\Gamma(u) = \{v \in V|(u, v) \in E\}$ is set of friends of u computed on E. The user behavior B can be regarded as two parts, namely content and interaction, which respectively represent open social activities, such as uploading or sharing logs, photos etc., and interactive actions, such as like, visit, comment, retweet etc., that target a specific user.

PROBLEM 1: Given a social network $G\langle$V,E,A,B$\rangle$, a user $u \in V$, an integer k and a quantitative metric Θ, the user link inference problem (**LIP** for short) is to find a set of k users $P_u^k = \{v_i|v_i \in V, i \in [1, k], (u, v_i) \notin E\}$ such that $\forall v_i \in P_u^k$ and $\forall v' \in V \cap v' \notin P_u^k$, $\Theta(u, v_i) > \Theta(u, v')$ holds.

The **LIP** problem is to find the top k probable friends for u against a metric. The key to this problem is to design an appropriate metric to capture the intrinsic factors that influence social links. In this paper, we take into account user attribute, social behavior and network structure to understand user social activities for link inference. Different metrics are created against these elements and denoted by Θ_a, Θ_b and Θ_s, respectively.

User preferences are considered from two aspects. The first aspect concerns user bias on friend attributes, which can be learned from current links and social activities. The quantitative biased weight on each attribute value is computed by means of population probabilities or entropy methods. The second aspect considers the bilateral wishes of users when evaluating a link, which outperforms the traditional approaches or an overall measurement.

To find the intrinsic characteristics that influence user social activities, we borrow from psychology and social science the idea of using latent factors to represent these information. Two social matrixes are then created against user content and interaction, and the latent factors are extracted by means of matrix decomposition. With respect to the network structure, we adopt some popular quantitative metrics for link inference, namely the common neighbor metric and the path metric.

To integrate those related metrics, we introduce two combination strategies: independent fusion and interdependent fusion. The former considers each metric independently and combines them by assigning different weights. Those weights are calculated by means of information entropy. The latter integrates multiple metrics into the existing network structure based models. Based on this principle, we propose some combinatorial metrics to calculate user closeness for link inference, which we discuss in details in the following sections.

IV. User Preference on Friend Attibutes

In real life, the concept of $social\ value$ is often used to measure the effect of a person being a friend to another. Obviously, different attributes will result in different social values being attributed to an individual. For example, some users would like to choose colleagues as friends while others may prefer as friends people with the same interest. So, we need to quantify the social value of attributes for different users so as to understand their preferences.

A. User Preference on Friend Attributes

A user's preferences with respect to friend attributes can be learned from the user's current links. Obviously, a larger population of friends indicates a higher social value for an attribute value. We thus introduce the **Population Based Attribute Importance** metric. For user u and attribute value a, the importance of a for u can be quantified as the proportion of u's friends holding a against the proportion holding a in the whole network. This reflects how much this attribute is important to u so that it can be used to infer a new link. For example, suppose that 50% of Tom's friends are from New York and that the population from New York in the entire network is only 5%. Then with respect to the geographic location attribute, a new user from New York holds higher probability of being Tom's friend than people from other areas. Formally, the attribute importance metric is given in equation 1, where $\Gamma(u)$ is the set of u's friends and I_u^a is the importance of attribute a.

$$I_u^a = \frac{|\{x \in \Gamma(u)|A_x.a = a\}|/|\Gamma(u)|}{|\{y \in V|A_y.a = a\}|/|V|} \tag{1}$$

The metric provides a direct quantification of the importance of an attribute value. However, it can not measure the comparative importance. For example, if the proportion of $u's$ friends holding a is larger than that of the whole network, but these friends are only a small part of u friends, then a is still not a distinct indicator in judging a link. So, it is necessary to further evaluate the information contained in each attribute value.

The **Entropy Based Quantification Importance** is then introduced to measure the comparative importance of each attribute value. In information theory, entropy is a measure of the uncertainty in a random variable [6]. In this context, the term refers to the confidence of choosing a user with the attribute value a as a friend for u. We compare two cases: the general probability of a person being a friend to u and the probability of a person with attribute value a being a friend of the same user. The more the uncertainty is reduced, the more information we obtain from this attribute for inferring $u's$ friends. Let p_{fu} represent the probability of a user being a friend of u in the social network G. When we have no knowledge about user attributes, p_{fu} is the ratio of $u's$ friends number against the user number in G, say $p_{f_u} = |\Gamma(u)|/|V|$. The probability of a user not being a friend of u is denoted by $p_{\bar{f}_u} = 1 - p_{f_u}$. So, the uncertainty $E(V)$ about whether a user being $u's$ friend or not is calculated by the information entropy. Formally, $E(V) = -(p_{f_u} log_2 p_{f_u} + p_{\bar{f}_u} log_2 p_{\bar{f}_u})$.

For attribute value a, the user set V can be partitioned into two parts: the set of users holding a and its complementary set, dented by V^a and $V^{\bar{a}}$, respectively. Based on this knowledge on a, the uncertainty is then computed as equation 2. The information gain $Gain(a)$ is the variation of these two information expectations as equation 3, which quantifies how the attribute value a reduces the uncertainty in link inference. So, it can be used as the metric for evaluating the importance of a for u.

$$E_a(V) = \frac{|V^a|}{|V|}E(V^a) + \frac{|V^{\bar{a}}|}{|V|}E(V^{\bar{a}}) \tag{2}$$

$$I_u^a = Gain(a) = E(V) - E_a(V) \tag{3}$$

Based on the attribute importance, $u's$ preference on attributes is represented as a vector $L_u^w = (I_u^1, \cdots, I_u^n)$, where n is the size of attribute value set, I_u^i, $i \in [1..n]$ is the importance of the corresponding attribute value for u.

B. Attribute Preference Based Link Metric

For a new user $v \in V$, an attribute vector $L_v = (l_v^1, \cdots, l_v^n)$ is created, where $l_v^i \in \{0, 1\}, i \in [1..n]$, $l_v^i = 1$ indicates that v has attribute value i, otherwise, $l_v^i = 0$. The probability of v being a friend to u is calculated as the similarity between u's preference and $v's$ qualification vectors. There are quite a few metrics to measure the similarity of two vectors. An example choice is the cosine measurement, which is sensitive to numerical values. Formally,

$$F(u \to v) = \frac{L_u^w \cdot L_v}{|L_u^w||L_v|} \tag{4}$$

Note that in the above equation, L_u^w reflects $u's$ preference on a friend attribute, while L_v evaluates whether $v's$ attribute is qualified for $u's$ preference. Considering that the creation of a link relies on bilateral wishes, we take into account mutual preferences of two users when evaluating a link. That is to say, not only u's preference on v is considered, but also the willingness of v is considered. So, the link metric between u and v is formalized as $\Theta_a(u, v)$:

$$\Theta_a(u, v) = \digamma(u \rightarrow v) + \digamma(v \rightarrow u) \tag{5}$$

V. Latent Factors in Social Behavior

According to psychology and social science, people often prefer to make friends among people with similar interests. Such observation has also been taken into account in previous approaches, such as approaches that use interest groups to predicate a link [11]. Unlike these approaches, we aim at finding latent factors that influence people social activities. The more similar the social motivations of two users are, the higher the probability that they will connect in the social network. This knowledge can be learned from user social activities. In this section, we analyze the two kinds of social activities: social contents and interaction.

A. Social Content Based Link Metric

Social contents refer to text, photo, post, comment, link to a blog, and web pages. We employ the information extraction method for social contents to create a user vector. For example, concerning text information, the Bag-of-Words model is used to extract a collection of words contained in text and the appearance of each word is independently counted. In the case of images, visual words are extracted by means of the Scale Invariant Feature Transform algorithm [19]. So for all social contents in a graph G, a dictionary is constructed, where each word is associated with a unique index. For each user, a vector is created according to one's social contents against the dictionary. For all users in G, we thus obtain a behavior matrix $B_c^{n \times m}$ (B_c for short hereafterwards), where $n = |V|$ and m is the number of words in the dictionary.

Overall, the behavior matrix is very sparse and it is thus difficult to analyze user correlations. To better understand these social activities, we introduce the latent factors to capture user social motivations by applying a matrix decomposition method on B_c [24], namely $B_c = N \times M^T$, where $N \in R^{n \times k}$, $M \in R^{m \times k}$ and k in the number of latent factors. N is the latent factor matrix associating each user with some latent factors. M is the behavior latent factor matrix representing the effect of latent factors on each behavior word. Our goal is to obtain an appropriate decomposition from which to determine a user potential social motivation. Let B_{ij} denote the cell with row i and column j of B_c. Let N_i and M_j denote the i^{th} row of matrix N and j^{th} row of matrix M, respectively. We define the conditional probability distribution over user behaviors as:

$$P(B_c|N, M, \sigma^2) = \prod_{i=1}^{n} \prod_{j=1}^{m} [D(B_{ij}|N_i M_j^T, \sigma^2)]^{I_{ij}} \tag{6}$$

where $D(x|\mu, \sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and variance σ^2, and I_{ij} is the indicator function that is equal to 1 if user i has the specific behavior word j and 0 otherwise. Zero-mean Gaussian priors are set for behavior feature vectors:

$$\begin{aligned} P(N|\sigma_N^2) &= \prod_{i=1}^{n} D(N_i|0, \sigma_N^2 I), \\ P(M|\sigma_M^2) &= \prod_{j=1}^{m} D(M_j|0, \sigma_M^2 I) \end{aligned} \tag{7}$$

Through the Bayesian inference, the log of the posterior distribution over user behavior is given by equation 8.

$$\begin{aligned} &lnP(N, M|B, \sigma_B^2, \sigma_N^2, \sigma_M^2) \propto P(B|N, M, \sigma_B^2)p(N|\sigma_N^2)p(M|\sigma_M^2) \\ &= ln \prod_{i=1}^{n} \prod_{j=1}^{m} P(B|N, M, \sigma^2) \times \prod_{i=1}^{n} P(N|\sigma_N^2) \times \prod_{j=1}^{m} P(M|\sigma_M^2) \\ &= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{m} I_{ij}(B_{ij} - N_i M_j^T)^2 - \frac{1}{2\sigma_N^2} \sum_{i=1}^{n} ||N_i||^2 \\ &\quad - \frac{1}{2\sigma_M^2} \sum_{j=1}^{m} ||M_j||^2 + C \end{aligned} \tag{8}$$

where N and M can be learned purely based on the user-behavior matrix using the gradient descent technique. Based on the idea that people prefer to make friends with other people similar to themselves, the content based link metric for users u and v can be calculated by equation 9, where N_u and N_v are the rows of user latent factor matric N.

$$\Theta_b(u, v) = \frac{N_u \cdot N_v}{|N_u||N_v|} \tag{9}$$

B. Interaction Behavior Based Link Metric

A user interaction activity targets a specific user, such as $visit$, $support$, $comment$, and $retweet$. For different users, the same interaction activity may contain different intimacy information. For example, for an active user with many social activities, a single $visit$ to another user does not indicate a high intimacy. Comparatively, for an inactive user with only few social activities, this $visit$ probably indicates a high intimacy. So interaction behaviors should be assigned biased weights for different users.

Let k denote the number of interaction behavior types. For user u, we create his/her interaction matrix as $B_l(u) \in R^{n \times k}$, where n is the number of users in the network. Let B_v denote the v^{th} row of $B_l(u)$ which shows $u's$ interaction behaviors towards user v. Each element B_{vi} refers to the number of interaction i towards v. To compute the importance of behavior type i of user u, we adopt methods similar to the ones for attributes. Let $E(V)$ denote the information entropy when we do not have any information about behavior i and $E_i(V)$ denote the information entropy when we know such information of i. So the social behavior weight I_u^i for user u with respect to behavior i is calculated by $I_u^i = Gain(i) = E(V) - E_i(V)$. The effort of all interactions from u to v is defined by equation 10, which can be regarded as the evaluation on how the amount of influence from u to v in a connecting path.

$$\mathbb{I}_{(u \rightarrow v)} = \sum_{i=1}^{k} B_{vi} \times I_u^i \tag{10}$$

Consider the bidirectional influence, the behavior matrix $B_l(v)$ for user v is created and $\mathbb{I}_{(v \to u)}$ is calculated. So the closeness of u and v on interaction behaviors can be calculated as follow:

$$\Theta_l(u, v) = \mathbb{I}_{(u \to v)} + \mathbb{I}_{(v \to u)} \tag{11}$$

VI. Link Metric Fusion

In the above sections, we have introduced several link metrics based on different knowledge on user attributes, social activities, and topologies. Since any single element may not be completely reliable, we integrate them for link inference. In this section, we introduce two combination strategies: independent fusion and interdependent fusion. The former considers each metric independently and then combines their results. The latter integrates multiple elements with network structure based models, such as Common Friends or Random Walk.

A. Independent Fusion

In the independent fusion, the considered metrics are combined together by assigned weights. For users u and v, let $\Theta_a(u, v)$, $\Theta_b(u, v)$ and $\Theta_l(u, v)$ denote the user attribute metric, social content metric, and interaction behavior metric for link inference, respectively. We also take into account the network structure based measurement, denoted by $\Theta_g(u, v)$. So, a fusion link metric can be expressed as equation 12, where w_i, $i \in [1..4]$ represents the weight of each factor.

$$\begin{aligned}\Theta_{IF}(u, v) = w_1 \cdot \Theta_g(u, v) + w_2 \cdot \Theta_a(u, v) \\ + w_3 \cdot \Theta_b(u, v) + w_4 \cdot \Theta_l(u, v)\end{aligned} \tag{12}$$

To assign a reasonable weight to each metric, we need to further understand the relative differences among these metrics by quantifying how much they reveal about a user preference with respect to choosing friends. The basic idea is to analyze the information that each metric contains. Since entropy is the quantification of the uncertainty in a probability distribution over several elements, it is a natural choice for representing this information. Considering that the value domain of a metric is the set of real numbers, we discretize the domain into several intervals $T_i, i \in [1, t]$, $t \in N^+$, and counting the occurrences $\Theta_x(u, v) \in T_i$ as $f(T_i)$, where Θ_x represents an alternative metric. The probability over T_i is then computed as $p_i = f(T_i)/\Sigma_{k=1}^{t} f(T_k)$. So the entropy of metric Θ_x is $E_{\Theta_x} = -\sum_{i=1}^{t} p_i log(p_i)$. A larger entropy means more uncertainty and thus we can obtain less information from the corresponding link metric. So, we adopt the inverse of entropy $w_i = \frac{1}{E_{\Theta_x}}$ as the weight in equation 12.

An obvious advantage of the independent fusion is that it is easy to combine several influential elements. However, sometimes, some intrinsic correlations may exist in these metrics, which are ignored in the independent fusion. To address this consideration, we propose the interdependent fusion mode in the following subsection.

TABLE I
INTERDEPENDENT FUSION OF DIFFERENT METRICS

	Attribute	Social content	Interaction
Common friend	A-CF	B-CF	L-CF
Random walk	A-RW	B-RW	L-RW
Paths	A-P	B-P	L-P

B. Interdependent Fusion

The interdependent fusion mode tightly couples multiple metrics together by integrating the influential factors with the network structure based models, such as the Common Friends and Random Walk models. Some possible combination methods are given in table I. For example, the Attribute and Common Neighbors based Metric ($A - CF$ for short) integrates the attribute metric with the common neighbor based measurement. We now discuss some representative combination metrics; other combinations can be obtained by using the same approach. In the following discussion, θ_x denotes a link metric, such as $\Theta_a(u, v)$, $\Theta_b(u, v)$ and $\Theta_l(u, v)$, or the combination of several metrics.

a) Common Friend Based Combination Metric: Let $\Gamma(u)$ denote the set of neighbors of user u in the social network G. The network structure based link inference shows that having a large number of common friends is an indication of the existence of a link. A direct implementation of this idea is to define the closeness of two users u and v as the number of their common neighbors, $\Theta_g(u, v) = |\Gamma(u) \bigcap \Gamma(u)|$, or to use the normalized Jaccard Coefficient, $\Theta_g(u, v) = (\Gamma(u) \bigcap \Gamma(v))/(\Gamma(u) \bigcup \Gamma(v))$. Obviously, the actual method does not take into account the different influences of friends. As we know from real life, one friend may have a higher influence than other friends in creating a new link. Such difference can be measured by a user closeness metric, such as $\Theta_a(u, v)$, $\Theta_b(u, v)$, or $\Theta_l(u, v)$. The take into account the weights of friends, we introduce the tight combination with common friends metric. For users u and v, let Z represent their common friend set, say $Z = \Gamma(u) \bigcap \Gamma(v)$. For each user $z \in Z$, we separately calculate his closeness with u and v. The harmonic average of their closenesses can be used as the diffusion effectiveness from z to u and v. So, the link metric between u and v is calculated as follow:

$$\Theta_{DF}(u, v) = \sum_{z \in Z} \frac{2 * \Theta_x(z, u) * \Theta_x(z, v)}{\Theta_x(z, u) + \Theta_x(z, v)} \tag{13}$$

b) Path Based Combination Metric: The common friend method only considers one-step connections between users without considering multi-distance connections, which however have been shown to affect link prediction [10], [21]. The more paths exist and the shorter each path, the higher the probability of a link. For example, the Katz' path measurement directly sums over this collection of paths [10], exponentially suppressed by path length so as to count short paths more heavily. Let $Katz\beta(u, v) = \Sigma_{\ell}^{\infty} \beta^{\ell} \cdot |paths_{u,v}^{\langle \ell \rangle}|$, where β is a constant used to limit the impact of long paths, $paths_{u,v}^{\langle \ell \rangle} = \{\text{paths of length } \ell \text{ from u to v}\}$, and $\ell >= 2$. Two variants of this measure are defined: (1) unweighted $paths_{u,v}^{\langle\ell\rangle}$; it is equal to 1 if and only if there is a path between u and v with size l and 0 otherwise; and (2) weighted $paths_{u,v}^{\langle\ell\rangle}$; it is equal to the number of paths of length ℓ between u and v.

In this metric, each link is considered the same, which therefore does not account for knowledge about the different influence of links on paths. For example, user closeness can be regarded as a link influence. Concerning user preferences, the proposed link metric $\theta_x(u,v)$ should be tightly combined with the weighted $Katz\beta$ method. Formally,

$$\Theta_{DF}(u,v) = \Sigma_{\ell}^{\infty} \beta^{\ell} * (\sum_{i=1}^{\ell-1} |\Theta_x(u_i, u_{i+1})|) \tag{14}$$

c) Random Walk based Combination Metric: Random walk is a network structure based model widely used in link prediction. It randomizes user relationships in a social network and the key parameter is this point-to-point access probability. In order to predict the friends of user u, a random walk is started from u and on each step we have a restart probability α to decide whether continuing the walk in the whole network or restarting from this new node. If the decision is to continue the walk, we should calculate the probability of a random walk to next node.

The traditional random walk algorithm is mostly used according to the number of user neighbors. To take into account the different influence of neighbors, we integrate the proposed link metrics Θ_x into the random walk method by replacing each link weight with user closeness. Then the probability of a random walk is the standardization of the corresponding weights. Suppose we are at node v, for each friend $f \in \Gamma(v)$, the probability p_{vf} of a random walk from v to f is computed as:

$$p_{vf} = \frac{\Theta_x(v,f)}{\sum_{f \in \Gamma(v)} \Theta_x(v,f)} \tag{15}$$

After a few iteration of the random walk, the access probability of each user node will converge to a number, denoted by Θ_{DF}, which can be used as the link metric.

VII. Experiments And Result Analysis

A. Dataset and Evaluation Metrics

In our experiment, we use two real databases. One is the most popular on-line social network in Slovakia, Pokec [26], which has been available for more than 10 years. This social network connects more than 1.6 million people. This dataset contains anonymous user data of the whole network. Another is Tencent Weibo [3], which has been added thousands of new users each day to the existing billions of active users, since its launch in April 2010. The dataset contains anonymous users for some parts of the network, including the links between users, user profile information which is anonymized into numbers, the user tweet and retweet data which is anonymized into keywords, and parts of the user interaction data. The detailed statistics are showed in Table II. The reason for choosing these datasets is that they contain more user information on attributes and interaction behaviors, while other public datasets only contain social network structures or some user attributes without user behaviors.

TABLE II
Data Sets

Dateset	#Nodes	#Edges	Attributes	Behaviors
Pokec	1632803	30622564	gender,age,hobby	NULL
Tecent Weibo	1379738	50655143	birth year,gender tweets, tag-ID	tweet,retweet comment

In both datasets the considered attributes include numeric attributes, like age, and categorical attributes, like gender. We preprocessed on the attributes. For numerical data, we discretize them into several intervals and each interval is treated as an attribute value. For enumerable data, each constant is treated as an attribute value. Then for each user, an attribute vector is created against his/her values. In Tencent Weibo, the behaviors are anonymous keywords which are extracted from tweets, retweets, and comments. Since the number of keywords is very large, we selected the most commonly used keywords. Considering that some users are without common friends or attributes and thus are not appropriate for the comparing our approach with other approaches, they were removed from the datasets. The training data is chosen as 80% of the entire dataset.

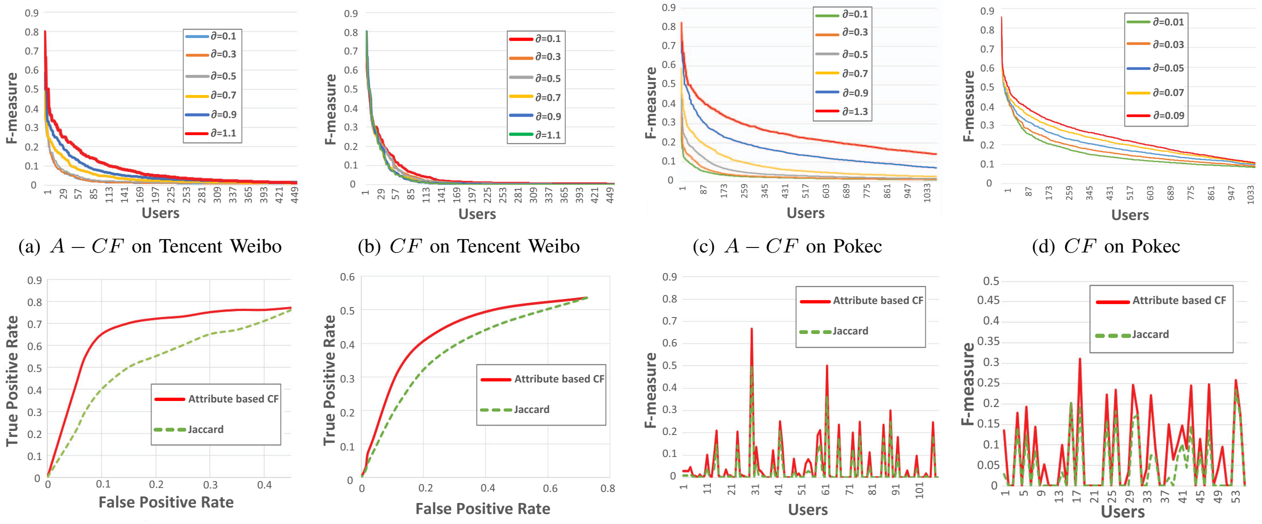
We adopt some commonly used evaluation metrics. For a predefined threshold, the results of link inference are labeled as either positive (P) or negative (N). The true positive (TP) case occurs when both the prediction outcome and the actual value are P. If the outcome is P and the actual value is N, this is a false positive (FP) case. Similarly, we have true negative (TN) and false negative (FN). Then the precision is $P_r = TP/(TP+FP)$ and the recall rate is $R_r = TP/(TP+FN)$. The metric F-measure is the harmonic average of P_r and R_r, $F = (2 * P_r * R_r)/(P_r + R_r)$. The ROC is a curve whose x-axis is $FPR = FP/(FP + TN)$ and y-axis is $TPR = TP/(TP + FN)$. The surface area under the curve is the accuracy of link reference.

B. Experimental Results

We verify the proposed link metrics from several aspects. First, we evaluate how the parameters and threshold affect the results of link inference. Then we compare the proposed different combination metrics with other methods.

We first compare the combined attribute and common friend fusion method $(A - CF)$ against the network structure based metric on two datasets. The results in Fig.1 (a)-(d) show the F-measure for different threshold values ∂ for each user on the x-axis, which is used to assess whether the inferred link (u,v) exists based on $\Theta(u,v) > \partial$. We can see that our method has better performance for increasing values of ∂ and reaches a stable value after $\partial = 1.3$ on $Pokec$ and $\partial = 1.1$ on $Tencent\ Weibo$, respectively. Comparatively, the performance of the common friend method works well for decreasing values of ∂ on $Tencent\ Weibo$, while it is almost not affected by

[3] www.kddcup2012.org/c/kddcup2012-track1/data



(a) $A-CF$ on Tencent Weibo (b) CF on Tencent Weibo (c) $A-CF$ on Pokec (d) CF on Pokec

(e) ROC on Tecent Weibo (f) ROC on Pokec (g) F on Tecent Weibo (h) F on Pokec

Fig. 1. Comparison on Attribute Based Fusion Methods with Traditional Methods

∂ on $Pokec$. We also consider the ROC curve and show the results in Fig.1 (e)-(f). Overall, our proposed metric is more accurate in evaluating links than the previous methods and has better performance than the previous methods. Since ROC evaluates a method only at the average level for all users, we also calculate the F-measure on each user. The results in Fig.1 (g)-(h) show that our method is better than the previous methods on most users.

Then we compare the attribute and random walk fusion method $(A-RW)$ and show the results in Fig.2. Since the random walk algorithm highly relies on the restart probability α, we evaluate the accuracy on different α settings. The results in Fig.2(a)-(d) show that both our method and the random walk method have better performance for increasing values of α on both datasets, and reach their best performance when $\alpha = 0.8$. The results on the experiments in Fig.2(e)-(h) show that our method outperforms the previous methods on both measurements, that is, F-measure and accuracy.

Finally, we evaluate the behavior and network structure based fusion metrics ($B-CF$ and $B-RW$). In order to fairly justify the behavior based method, we randomly select 100 users who are associated with friends and behavior information in $Tencent\ Weibo$. We first compare the $B-CF$ metric with the common friend method. The results, reported in Fig.3(a), show that our method, represented by the red curve, has higher values of the F-measure in most cases than the common friend method, represented by the green dashed curve. Then we compare the fusion metric $(B-RW)$ with the traditional random walk method. The results, reported in Fig.3(b), show that our method is better than the random walk method in most cases. Besides, the performances of both methods grow with the increase of restart times and reach a stable value after it is 14000. To be mentioned here, although our improvement seems not distinctly large with respect to accuracy, when taking into account the size of link set in predication, the number of inferred links highly outperforms other methods.

VIII. Conclusions

In this paper, we investigate the link inference problem in social services, which has significance in economy, social security and other areas. To gain a deeper understanding of user behavior, we introduce the concept of latent factor to capture the intrinsic correlations between social purpose and behavior. User preferences are considered in link inference beyond the traditional methods or overall measurements. To semantically combine these measurements together for link inference, we propose the independent fusion and interdependent fusion methods. Experimental results on real datasets show that our approach outperforms previous approaches.

Link inference is actually a kind of recommendation. Our methods can be also applied to the selection of web services or service recommendation. As part of future work, we will consider the evolution of user social preference. In practice, user preferences may change according to one life stages, which should be taken into account for link inference.

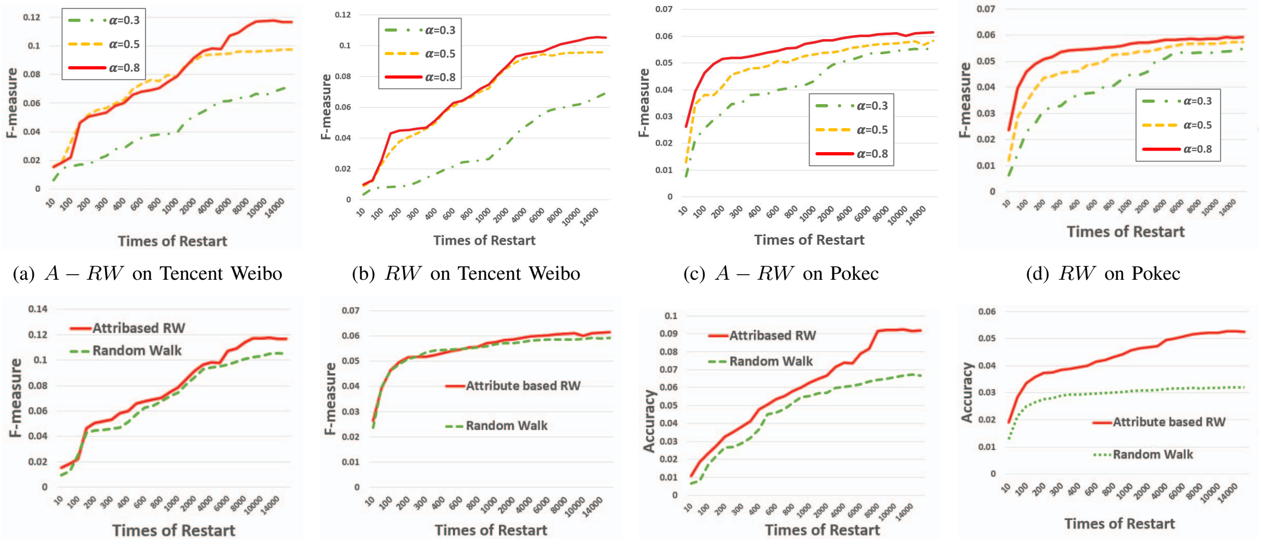
Acknowledgment

This work is supported by the NSF China (61173140), SAICT Experts Program, Independent Innovation & Achievements Transformation Program (2014ZZCX03301) and Science & Technology Development Program of Shandong Province (2014GGX101046).

References

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

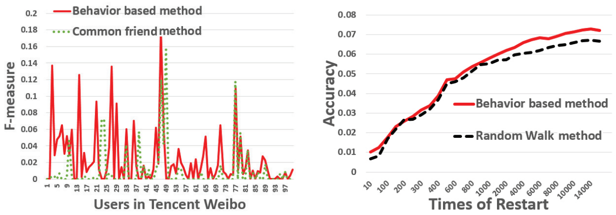
[2] C. G. Akcora, B. Carminati, and E. Ferrari. User similarities on social networks. *Social Network Analysis and Mining*, 3(3):475–495, 2013.



(a) $A - RW$ on Tencent Weibo (b) RW on Tencent Weibo (c) $A - RW$ on Pokec (d) RW on Pokec

(e) F on Tencent Weibo (f) F on Pokec (g) P_r on Tencent Weibo (h) P_r on Pokec

Fig. 2. Comparison on Attribute and Random Walk Fusion Methods



(a) Common Friend (b) Random Walk

Fig. 3. Compare behavior based methods

[3] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM06: Workshop on Link Analysis, Counterterrorism and Security*, 2006.

[4] H. Chen, W.-S. Ku, H. Wang, L. Tang, and M.-T. Sun. Linkprobe: Probabilistic inference on large-scale social networks. In *Proceedings of the IEEE 29th ICDE*, pages 290–301. IEEE, 2013.

[5] N. Z. Gong, A. Talwalkar, L. Mackey, L. Huang, E. C. R. Shin, E. Stefanov, E. R. Shi, and D. Song. Joint link prediction and attribute inference using a social-attribute network. *ACM TIST*, 5(2):27, 2014.

[6] S. Ihara. *Information theory for continuous systems*, volume 2. World Scientific, 1993.

[7] M. Jiang, Y. Chen, and L. Chen. Link prediction in networks with nodes attributes by similarity propagation. *arXiv:1502.04380*, 2015.

[8] T. Jin, T. Xu, E. Chen, Q. Liu, H. Ma, J. Lv, and G. Hu. Random walk with pre-filtering for social link prediction. In *Proceedings of the 9th CIS, 2013.*, pages 139–143. IEEE, 2013.

[9] H. Kashima and N. Abe. A parameterized probabilistic model of network evolution for supervised link prediction. In *Proceedings of the Sixth International Conference on Data Mining*, pages 340–349. IEEE, 2006.

[10] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[11] V. Leroy, B. B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD*, pages 393–402. ACM, 2010.

[12] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th WWW*, pages 641–650. ACM, 2010.

[13] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *JASIST*, 58(7):1019–1031, 2007.

[14] F. Liu, B. Liu, C. Sun, M. Liu, and X. Wang. Deep belief network-based approaches for link prediction in signed social networks. *Entropy*, 17(4):2140–2169, 2015.

[15] F. Liu, B. Liu, C. Sun, M. Liu, and X. Wang. Multimodal learning based approaches for link prediction in social networks. In *Natural Language Processing and Chinese Computing*, pages 123–133. Springer, 2015.

[16] F. Liu, B. Liu, X. Wang, M. Liu, and B. Wang. Features for link prediction in social networks: A comprehensive study. In *Proceedings of the 2012 IEEE SMC*, pages 1706–1711. IEEE, 2012.

[17] Q. Liu, J. Li, Z. Xie, and P. Zhang. An improvement of link prediction by combining local information and betweenness. In *Proceedings of the 2015 11th ICNC*, pages 456–461. IEEE, 2015.

[18] Z. Liu, Q.-M. Zhang, L. Lü, and T. Zhou. Link prediction in complex networks: a local naïve bayes model. *EPL*, 96(4):48007, 2011.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[20] M. E. Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102, 2001.

[21] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos. Fast and accurate link prediction in social networking systems. *Journal of Systems and Software*, 85(9):2119–2132, 2012.

[22] B. Qiu, Q. He, and J. Yen. Evolution of node behavior in link prediction. In *AAAI*, 2011.

[23] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM WSDM*, pages 723–732. ACM, 2012.

[24] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. Citeseer, 2011.

[25] Y. Sun, J. Han, J. Gao, and Y. Yu. itopicmodel: Information network-integrated topic modeling. In *Proceedings of the Ninth IEEE International Conference on Data Mining, 2009.*, pages 493–502. IEEE, 2009.

[26] L. Takac and M. Zabovsky. Data analysis in public social networks. In *International Scientific Conference and International Workshop Present Day Trends of Innovations*, pages 1–6, 2012.

[27] C. Wang, V. Satuluri, and S. Parthasarathy. Local probabilistic models for link prediction. In *Proceedings of the Seventh IEEE International Conference on Data Mining, 2007.*, pages 322–331. IEEE, 2007.

[28] P. Wang, B. Xu, Y. Wu, and X. Zhou. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, 58(1):1–38, 2015.

[29] X. Yang, H. Steck, and Y. Liu. Circle-based recommendation in online social networks. In *Proceedings of the 18th ACM SIGKDD*, pages 1267–1275. ACM, 2012.

[30] B. Zhu and Y. Xia. An information-theoretic model for link prediction in complex networks. *Scientific reports*, 5, 2015.