Learning Latent Friendship Propagation Networks with Interest Awareness for Link Prediction*

Jun Zhang[1,2,3,4] Chaokun Wang[2,3,4] Philip S. Yu[5] Jianmin Wang[2,3,4]
[1] Department of Computer Science and Technology, Tsinghua University
[2] School of Software, Tsinghua University
[3] Tsinghua National Laboratory for Information Science and Technology
[4] Key Laboratory for Information System Security, Ministry of Education, P. R. China
[5] Department of Computer Science, University of Illinois at Chicago
zhang-jun10@mails.thu.edu.cn, chaokun@tsinghua.edu.cn, psyu@uic.edu, jimwang@tsinghua.edu.cn

ABSTRACT

It's well known that the transitivity of friendship is a popular sociological principle in social networks. However, it's still unknown that to what extent people's friend-making behaviors follow this principle and to what extent it can benefit the link prediction task.

In this paper, we try to adopt this sociological principle to explain the evolution of networks and study the latent friendship propagation. Unlike traditional link prediction approaches, we model link formation as results of individuals' friend-making behaviors combined with personal interests. We propose the Latent Friendship Propagation Network (LFPN), which depicts the evolution progress of one's egocentric network and reveals future growth potentials driven by the transitivity of friendship based on personal interests. We model individuals' social behaviors using the Latent Friendship Propagation Model (LFPM), a probabilistic generative model from which the LFPN can be learned effectively. To evaluate the power of the friendship propagation in link prediction, we design LFPN-RW which models the friend-making behavior as a random walk upon the LFPN naturally and captures the co-influence effect of the friend circles as well as personal interests to provide more accurate prediction.

Experimental results on real-world datasets show that LFPN-RW outperforms the state-of-the-art approaches. This convinces that the transitivity of friendship actually plays important roles in the evolution of social networks.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—*Data mining*; J.4 [**SOCIAL AND BEHAVIORAL SCIENCES**]: Sociology

Keywords

Link Prediction; Social Networks; Transitivity of Friendship; Friendship Propagation; Interest Awareness

*Corresponding authors: Chaokun Wang and Jianmin Wang.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

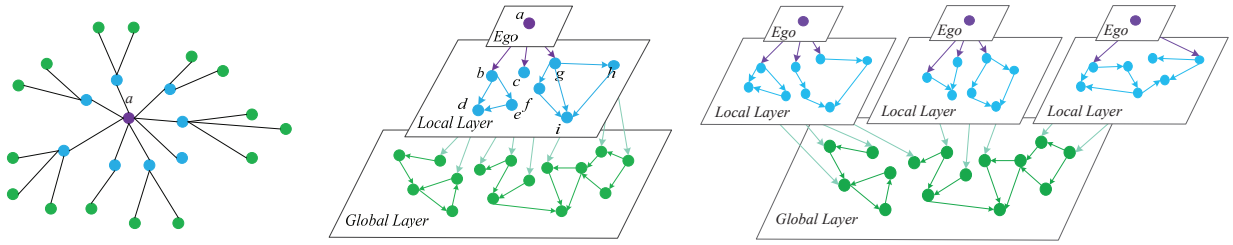
1. INTRODUCTION

The evolution of social networks has attracted considerable attention recently. Since the seminal work of Liben-Nowell and Kleinberg [18], it has been formulated as the link prediction task which studies the formation and variation of the links in the network. Both unsupervised and supervised approaches have been proposed, based on the current observations of the network, utilizing the structural or vertex-intrinsic features [19][9][20][1]. Most approaches treat the network as an evolving graph and each individual as a general vertex in the graph, while ignoring the initiative of individuals as active participants in the evolution of the social network.

From another point of view, people's social behaviors have been studied by sociologists long ago [28]. They have found many interesting and popular phenomena [23][21][26][8], which are also verified in online social networks by computer scientists [12][5][24][11][27]. Among these phenomena is triadic closure [26][8], which states that two persons with common friends will likely become friends in the future. The sociological principle under this phenomenon is the transitivity of friendship [26][15]. Although researchers have found that most of new links close an open triangle [16], not all open triangles will be closed with equal probabilities. It's still unknown how this principle works in real world, and to what extent it can promote the evolution of social networks and furthermore benefit the link prediction task.

In this paper, we study the evolution of social network based on one of the well-known sociological principles, i.e. the transitivity of friendship. We address the above problems by positing that there is some underlying unknown network over which friendships propagate, and the evolution of social networks can be regarded as results of the friendship propagation based on common user interests. We formulate such network as the Latent Friendship Propagation Network (LFPN), which depicts the growth process of one's ego-centric social network driven by the friendship propagation via the co-influence effect of the friend circles with common user interests, and also captures the trend of friendship propagation on the whole network. In LFPN, we assume one, denoted by the *ego*, expands her social circles mainly via her existing friends, who act as *intermediaries* and introduce new friends to the ego explicitly or implicitly based on common interests.

We demonstrate the proposed LFPN with a toy example. Fig. 1(a) shows an egocentric network with a as the ego, which we mark as purple. The blue vertices are a's friends; the green ones show the friends of a's friends. The connections among the blue and green vertices are omitted for brevity. Given the plain structure of an egocentric network, we can hardly observe the implicit intrinsic



(a) Egocentric network of a (b) a's partial view of the LFPN (c) More complete view of the LFPN

Figure 1: An LFPN is composed of three layers: the ego layers, the local layers for each ego, and the global layer for the whole network. In this figure, the purple vertex a is the ego, the blue vertices are the friends of the ego and the green ones show other individuals in the network.

structures and relations within it. The LFPN shows a 3-layer hierarchical network for the original network. Fig. 1(b) shows the partial view of the ego a. While the first layer is the ego herself, the second layer, the *local layer*, consists of all her current friends and the directed edges as representation of the friendship propagation traces along which a created relationships with each one. In this figure we can see that a made friends with b directly, and via b she became friends of e, and furthermore she created friendship with d via b and e. Through the *local layer* as intermediaries a can make more new friends in the *global layer* in which the edges among individuals represent the potential friendship propagation directions in the whole network. While the *local layer* is defined for each ego and visualizes the growth of one's friend circles by showing the intermediaries for each of her friends, the *global layer* is inferred based upon all the *local layers* and defined for the whole network. A more complete LFPN with multiple egos is shown in Fig. 1(c).

However, learning the LFPN is challenging because the network over which friendship propagation takes place is usually unknown and unobserved. Thus, the first challenge is, *how we can leverage one's social behaviors to infer all the potential intermediaries and furthermore learn the structure of the friendship propagation network?*

While traditional approaches for link prediction are usually based on the topological features, LFPN provides a hierarchical view of one's friend circles from which we can observe the evolution of networks from the perspective of the ego, who drives this evolution progress. This leaves us the second challenge: *to what extent the transitivity of friendship influences the network evolution, whether there is a co-influence effect of the friend circle, and how they can be utilized to predict future shapes of the network?*

Along with the transitivity of friendship, another factor which affects one's social behaviors is the personal interests. In this study, we're also interested in the challenge that *how the personal interests, along with the friendship propagation, affect the creation of friendships?*

1.1 Contributions

To address these challenges we study how the LFPN can be inferred based on behavior modeling, and further utilized for predicting or recommending new links. To the best of our knowledge, this is the first time that the transitivity of friendship is deeply investigated and the link formation is studied based on the behavior modeling. The main contributions of this paper can be summarized as follows.

Firstly, we formulate the **LFPN**, which depicts the evolution traces of one's egocentric social network and potential friendship propagation traces on the whole network based on the transitivity of friendship with interest awareness.

We design **LFPM** (Latent Friendship Propagation Model), a probabilistic generative model for the friend-making behaviors of individuals in social networks. LFPM models the interest-aware friendship propagation driven by the transitivity of friendship. Given the time-stamped individuals' social behavior histories, we can learn the hidden LFPN based on LFPM effectively. It should be noted that both the friendship propagation and personal interests are inferred from the social behaviors only, and no other evidences are included.

Furthermore, we put forward a new method for link prediction utilizing the transitivity of friendship and personal interests based on the LFPN. We simulate the link creation using a random walk of an ego on the two layers of the LFPN. Firstly, she seeks for a preferred intermediary in her *local layer* of the LFPN, and via this intermediary she goes to the *global layer* where she makes new friends. We name this process as **LFPN-RW** (Random Walk upon LFPN), in which the random walk is guided by the friendship propagation based on the co-influence effect of friend circles and personal interests.

We conduct extensive experiments on real-world datasets. Experimental results show that the inferred LFPN actually contains important knowledge about the network evolution and can benefit the link prediction significantly.

1.2 Roadmap

The rest of this paper is organized as follows. We formulate the LFPN in Section 2. Methods for inferring the LFPN are developed in Section 3. Afterwards we present LFPN-RW, the new approach for link prediction upon LFPN in Section 4. In Section 5, we discuss the experimental results. We review related work in Section 6 and conclude this study in Section 7.

2. LFPN DEFINITION

In this section, we introduce the concepts related to LFPN. As shown in Fig. 1(c), the LFPN $\mathcal{G}$ is composed of three layers: the ego, the local layer and the global layer. In the following we first define the local layer and global layer respectively, and then give the definition of the LFPN.

Let $\mathbb{G} = \{G_1, G_2, \cdots, G_T\}$ denote an evolving social network where $G_t = (V_t, E_t)$ is a directed graph denoting the snapshot of the network at time t. Each vertex in V_t represents an individual in the network, and the edges in E_t indicate the friendships among them. ΔG_t denotes the new vertices and edges appearing during the period t (i.e. the time from $t-1$ to t), and thus we have $G_t = G_{t-1} \cup \Delta G_t$. We'll use G to denote any snapshot of the evolving $\mathbb{G}$ for simplicity. Furthermore, for each snapshot G_t we have a corresponding interaction set $X_t \subset V_t \times V_t$ which denotes

the interactions among individuals in t. The interactions include co-authoring a paper in scholar collaboration network, leaving a message on Facebook, etc.

It should be noted that although the network G is directed here, the problem discussed in this paper is also applicable to undirected networks which can be transformed to directed networks.

In this paper, we study how E grows in social networks, i.e. the formation of friendships. Based on the transitivity of friendship, when two individuals become friends, we assume their common friends act as the role of "*intermediaries*".

Definition 1. **Latent Friendship Propagation Triple** (*LFP Triple*). For three vertices u, z and v in G, (u, z, v) is named as an **LFP Triple** if u makes friends with v via z. u is called the initiator of this tuple and z is the intermediary.

Each LFP Triple (u, z, v) is associated with an **intermediation probability** $p(z|\langle u, v\rangle)$ which denotes the probability that z has acted as the intermediary for u to makes friends with v.

We use $F_t(u)$ to denote the set of friends of u at time t, and $F_t(u, v)$ to denote the set of common friends of u and v at time t. As each common friend in $F_t(u, v)$ may act as the intermediary, the intermediation probability satisfies:

$$\sum_{z \in F_t(u,v)} p(z|\langle u, v\rangle) = 1. \quad (1)$$

If u doesn't have any common friends with v, we assume u makes friends with v directly.

Definition 2. The **local layer** of ego u on an LFPN $\mathcal{G}$, denoted as $\mathcal{G}^{L(u)}$, consists of all the friends of u. For each LFP Triple (u, z, v) initiated by u there is a corresponding *local LFP edge* $\langle z, v\rangle$ in $\mathcal{G}^{L(u)}$, whose weight $w^{L(u)}_{z,v} = p(z|\langle u, v\rangle)$.

The *local layer* of the LFPN is defined for each ego u. Each friend of the ego $u \in \mathcal{G}^{L(u)}$ can play the role of intermediary for u to make more new friends. We use the **intermediary preference probability** $p(u \rightarrowtail z)$ to measure u's preference for each friend z as the intermediary.

We define the cross-layer edges from the *ego* to the *local layer* as *intermediary preference edges*.

Definition 3. For each friend $z \in \mathcal{G}^{L(u)}$, there's an **intermediary preference edge** $\langle u, z\rangle$ pointing from the ego u to z in the local layer, whose weight $w^{E}_{u,z} = p(u \rightarrowtail z)$.

Definition 4. Each LFP Triple (u, z, v) derives an **LFP Pattern** $z \twoheadrightarrow v$, indicating that one's friendship with z is likely to propagate to v.

Each LFP Pattern is associated with a **friendship propagation probability** $p(z \twoheadrightarrow v)$ which denotes the extent to which z is willing to recommend v to her friends and the friends of z are likely to make friends with v in the future.

Definition 5. The **global layer** of an LFPN $\mathcal{G}$, denoted as $\mathcal{G}^{G}$, consists of all the individuals of the given social network. For each LFP Pattern $z \twoheadrightarrow v$ there's a corresponding *global LFP edge* $\langle z, v\rangle$ in $\mathcal{G}^{G}$, whose weight $w^{G}_{z,v} = p(z \twoheadrightarrow v)$.

The *global layer* is defined for the whole network based on the LFP Patterns. It can be interpreted as the aggregation of all the *local layers*. Each vertex has its direct projection on the *global layer* so we don't define the cross-layer from the *local* to *global* layer again.

Now we can give the definition of LFPN.

Table 1: The notations for the LFPM

U, K	# of individuals and topics
N_u, V_u	# of new friends and interactions of u
$F(u)$	The current friend set of individual u
$Y(u)$	The new friend set of individual u
$X(u)$	The interaction collection of individual u
θ_u	The intermediary preference distribution of u
ϕ_u	The friendship strength distribution of u
χ_u	The interest preference distribution of u
ψ_c	The reputation distribution of topic c
$z_{u,y}$	The intermediary assignment of the friendship $u \rightarrow y$
$\alpha, \beta, \delta, \gamma$	The prior parameters for θ, ϕ, χ, ψ, respectively
$\Omega^{(cz)}_{c_i,z_i}$	# that z_i has acted an intermediary in topic c_i
$\Omega^{(cy)}_{c_i,y_i}$	# that y_i has been recommended and accepted as a new friend in topic c_i
$\Omega^{(uc)}_{u_i,c_i}$	# that u_i has selected friends in topic c_i
$\Omega^{(uz)}_{u_i,z_i}$	# that u_i has selected z_i as an intermediary
$\Omega^{(zy)}_{z_i,y_i}$	# that z_i has recommend y_i for others
$\Omega^{(uv)}_{u_i,y_i}$	# that u_i has interacted with y_i

Definition 6. The **Latent Friendship Propagation Network** (LFPN) $\mathcal{G}$ for a given social network G is a weighted 3-layer network consisting the *ego layers* $\{u\}$, the *local layers* for each ego $\{\mathcal{G}^{L(u)}\}$, and the *global layer* $\mathcal{G}^{G}$.

The *local layer* shows the inferred traces for the evolution history of one's current egocentric network, while the *global layer* indicates the possible traces for the friendship propagation on the whole social network in the future. The *global layer* can also be regarded as a friend-recommendation network, as in some sense it's the recommendation among people that promotes the friendship propagation.

3. LFPN INFERENCE

In this section, we develop effective algorithms for the inference of LFPN for the given social network. Before proceeding, we formulate our problem as follows:

LFPN Inference Problem. Given an evolving social network $\mathbb{G} = \{G_1, G_2, \cdots, G_t\}$ and the associated interaction sets $\mathbb{X} = \{X_1, X_2, \cdots, X_t\}$, the LFPN inference problem is to infer the LFPNs $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_t$, where $\mathcal{G}_t$ is the LFPN for G_t.

As each snapshot is a subgraph of any of later snapshots, each LFPN also contains the structures of the LFPN inferred from the earlier snapshots. However, the weights of the same edges are not necessarily equal because the network is evolving and the propagation patterns are also varying.

In the remainder of this section, we first introduce the LFPM model for learning the LFPN on a single snapshot, and then present the inference framework for LFPM. We discuss the inference of LFPN on an evolving network at last.

3.1 The LFPM Model

Each LFP Triple (u, z, v) implies the assumption that the friend-making behavior is a 2-step process: u selects z as an intermediary first and then via z she makes friends with v. Meanwhile this process is inevitably influenced by the interests of u. LFP Patterns are the aggregation of LFP Triples, and both of them are the basic elements for LFPN. Thus the preliminary problem for the LFPN inference is to find all the LFP Triples. In other words, given each friendship between u and v, we need to infer the intermediary z.

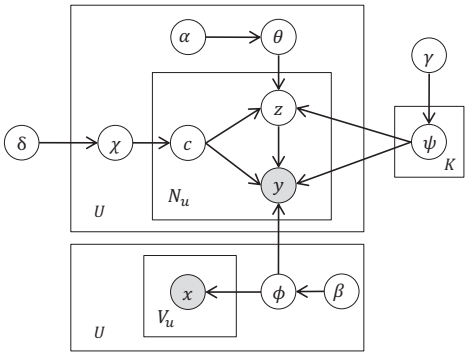


Figure 2: The graphical representation of LFPM

To model this process and infer the intermediaries of each friendship, we propose the Latent Friendship Propagation Model (LFPM), a probabilistic generative model for the social behaviors in social networks. In this paper we only consider the social behaviors like friend-making and interactions. LFPM models the generative process for individuals' social behaviors as follows:

1. Sample the number of individuals $U \sim \text{Poisson}(\epsilon)$.
2. For each of the K interest areas c, sample its *reputation distribution* $\psi_c \sim \text{Dir}(\gamma)$
3. For each individual u:
   (a) Sample her *friendship strength distribution* $\phi_u \sim \text{Dir}(\beta_u)$.
   (b) Sample the number of her interactions, $V_u \sim \text{Poisson}(\xi)$.
   (c) For each of the V_u interactions:
      i. Sample one of her existing friends, $x \sim \phi_u$, as another participant of the interaction
4. For each individual u:
   (a) Sample her *interest preference distribution* $\chi_u \sim \text{Dir}(\delta)$.
   (b) Sample her *intermediary preference distribution* $\theta_u \sim \text{Dir}(\alpha_u)$.
   (c) Sample the number of her new friends, $N_u \sim \text{Poisson}(\zeta)$.
   (d) For each of the N_u new friends:
      i. Sample an interest area $c \sim \chi_u$
      ii. Sample an intermediary $z \sim \theta_u \cdot \psi_c$
      iii. Sample a new friend $y \sim \phi_z \cdot \psi_c$

We draw the graphical representation of LFPM in Fig. 2. In LFPM, each interest area is modelled as a distribution over individuals. Each ego is modelled as a distribution of her intermediaries, i.e. the individuals that may make recommendations to her, and each intermediary as a distribution over the friends of this intermediary. Thus, the set of new friends of an ego can be modelled as the mixture of her intermediaries and interest areas. Specifically, the *intermediary preference distribution* θ describes through whom an ego prefers to make new friends, and *friendship strength distribution* ϕ models the closeness of an ego with each of her friends. Such closeness can be reflected by both the frequency of their interactions and the willingness of the ego to recommend her friends to others. We use the *interest preference distribution* χ to model one's interest preference in friend-making behaviors, and regard the *reputation distribution* ψ as a global reputation ranking of all individuals in each interest area. It's natural to assume that one prefers to make friends with someone with higher reputation in her preferred area. It should be noted that the interests here are learned only from the social behavior histories, rather than external profiles.

To improve the inference performance, we adopt heuristic pruning by constraining that for each individual her intermediary preference distribution and friendship strength distribution are only defined over her current friends, based on the assumption that only the current friends of ego u may act as u's intermediaries, and that an intermediary z can only recommend and interact with the current friends of z. This is different from traditional topic models like LDA[2], which defines the distributions over the whole space. Thus in our model, each individual maintains a unique sampling space for herself. Correspondingly, the prior distributions for θ and ϕ are also required to be defined on the corresponding pruned space for each individual. That's why the priors α and β need to be defined for each individual u and are placed within the big box labelled U in the graphical model shown in Fig. 2.

LFPM applies to a snapshot G_t and tries to infer the friendship propagation in the expansion of one's egocentric network in period t. For the social behaviors that happen in period t, the hidden distributions θ and ϕ of LFPM can be naturally interpreted as the intermediary preference probabilities w^E of each ego and friendship propagation probabilities w^G in the LFPN in period t, respectively. Each possible intermediary assignment for each friendship is an LFP Triple and the sampling probability can be regarded as the intermediation probabilities w^L. With the inferred LFP Triples and the weights w^E, w^L and w^G we can construct the LFPN $\mathcal{G}_t$.

3.2 Inference Framework for LFPM

Exact inference for LFPM is generally intractable. In this subsection we present the framework for LFPM inference using Gibbs sampling, a special case of Markov-Chain Monte Carlo (MCMC) simulation, which can emulate high-dimensional probability distributions by the stationary behavior of a Markov chain.

Firstly, the joint probability of all observed and unobserved data can be stated as follows (with the distributions integrated out):

$$\begin{aligned} &P(C,Z,X,Y;\alpha,\beta,\gamma,\delta) \\ &= (\prod_{u=1}^{U}\frac{B(\Omega_u^{(uz)}+\alpha_u^{(\theta)})}{B(\alpha_u^{(\theta)})})(\prod_{k=1}^{K}\frac{B(\Omega_k^{(cz)}+\Omega_k^{(cy)}+\gamma)}{B(\gamma)}) \\ &\cdot(\prod_{u=1}^{U}\frac{B(\Omega_u^{(uv)}+\Omega_u^{(zy)}+\beta_u^{(\phi)})}{B(\beta_u^{(\phi)})})(\prod_{u=1}^{U}\frac{B(\Omega_u^{(uc)}+\delta)}{B(\delta)}), \end{aligned} \tag{2}$$

where $B(\cdot)$ is the multinomial Beta function. For the details of other notations please refer to Table 1.

We infer the model by updating the estimation for unobserved variables c and z iteratively. In each iteration, we sample c and z for each new friendship i according to the conditional probability $p(c_i|C_{-i},Z,X,Y;\alpha,\beta,\gamma,\delta)$ and $p(z_i|C,Z_{-i},X,Y;\alpha,\beta,\gamma,\delta)$. Y and X are the new friendship and the interaction data, respectively. C and Z are estimations for the interest area and intermediary for each friendship, respectively. C_{-i} is the current estimations except c_i and Z_{-i} is that except z_i.

The sampling probabilities can be estimated by

$$\begin{aligned} p(c_i|C_{-i},Z,X,Y;\alpha,\beta,\gamma,\delta) &= \frac{P(C,Z,X,Y)}{P(C_{-i},Z,X,Y)} \\ &\propto \frac{(\Omega_{c_i,z_i}^{(cz)}+\Omega_{c_i,y_i}^{(cy)}+\gamma_{c_i,z_i}-1)}{\sum_{v\in F(u_i)}(\Omega_{c_i,v}^{(cz)}+\Omega_{c_i,v}^{(cy)}+\gamma_{c_i,v})-1} \\ &\cdot\frac{\Omega_{u_i,c_i}^{(uc)}+\delta_{u_i,c_i}-1}{\sum_{k=1}^{K}(\Omega_{u_i,k}^{(uc)}+\delta_{u_i,k}-1)}, \end{aligned} \tag{3}$$

$$\begin{aligned} p(z_i|C,Z_{-i},X,Y;\alpha,\beta,\gamma,\delta) &= \frac{P(C,Z,X,Y)}{P(C,Z_{-i},X,Y)} \\ &\propto \frac{\Omega_{u_i,z_i}^{(uz)}+\alpha_{u_i,z_i}^{(\theta)}-1}{\sum_{v\in F(u_i)}(\Omega_{u_i,v}^{(uz)}+\alpha_{u_i,v}^{(\theta)})-1} \\ &\cdot\frac{\Omega_{z_i,y_i}^{(uv)}+\Omega_{z_i,y_i}^{(zy)}+\beta_{z_i,y_i}^{(\phi)}-1}{\sum_{v\in F(u_i)}(\Omega_{z_i,v}^{(uv)}+\Omega_{z_i,v}^{(zy)}+\beta_{z_i,v}^{(\phi)})-1} \\ &\cdot\frac{\Omega_{c_i,z_i}^{(cz)}+\Omega_{c_i,z_i}^{(cy)}+\gamma_{c_i,z_i}-1}{\sum_{v\in F(u_i)}(\Omega_{c_i,v}^{(cz)}+\Omega_{c_i,v}^{(cy)}+\gamma_{c_i,v})-1}. \end{aligned} \tag{4}$$

Algorithm 1 LFPM Inference using Gibbs Sampling

Require:
The snapshot network G_{t-1} at time $t-1$
The increment of the network in period t, ΔG_t
The interaction set in period t, X_t
Ensure:
The hidden distributions θ, ϕ, χ, ψ
The sampling distributions P^c, P^z
1: Initialize F, Y using $G_{t-1}, \Delta G_t$, respectively
2: Assign interest areas for each new friendship randomly
3: Assign intermediaries for each new friendship randomly
4: **while** not finished **do**
5: **for** each individual u who has new friends **do**
6: **for** each new friend $v \in Y(u)$ **do**
7: Sample interest area using Eq. 3
8: Sample intermediary using Eq. 4
9: **end for**
10: **end for**
11: **end while**
12: Update distributions θ, ϕ, χ, ψ using Eq. 5 – 8
13: Record all the sampling distributions for topics and intermediaries in the last iteration with P^c and P^z
14: **return** $\theta, \phi, \chi, \psi, P^c, P^z$

Given the estimations for all unobserved data, we can estimate the hidden distributions:

$$\theta_{u_i,z_i} = \frac{\Omega^{(uz)}_{u_i,z_i} + \alpha^{(\theta)}_{u_i,z_i}}{\sum_{v \in F(u_i)} (\Omega^{(uz)}_{u_i,v} + \alpha^{(\theta)}_{u_i,v})}, \tag{5}$$

$$\phi_{z_i,y_i} = \frac{\Omega^{(uv)}_{z_i,y_i} + \Omega^{(zy)}_{z_i,y_i} + \beta^{(\phi)}_{z_i,y_i}}{\sum_{v \in F(u_i)} (\Omega^{(uv)}_{z_i,v} + \Omega^{(zy)}_{z_i,v} + \beta^{(\phi)}_{z_i,v})}, \tag{6}$$

$$\chi_{u_i,c_i} = \frac{\Omega^{(uc)}_{u_i,c_i} + \delta_{u_i,c_i}}{\sum_{k=1}^{K} (\Omega^{(uc)}_{u_i,k} + \delta_{u_i,k})}, \tag{7}$$

$$\psi_{c_i,u_i} = \frac{\Omega^{(cz)}_{c_i,u_i} + \Omega^{(cy)}_{c_i,u_i} + \gamma_{c_i,u_i}}{\sum_{u=1}^{U} (\Omega^{(cz)}_{c_i,u} + \Omega^{(cy)}_{c_i,u} + \gamma_{c_i,u})}. \tag{8}$$

Algorithm 1 presents the inference framework. Firstly, it initializes the friends set F before period t, and the new friendships Y that emerge in t. The interest area and intermediary for each new friendship are initialized randomly before the iteration. Next in Line 4–11 the interest and intermediary are updated for each new friendship using Eq. 3 and 4, iteratively. The iteration is repeated until converge or the count of iterations reaches a given threshold. The hidden distributions can be estimated using Eq. 5 - 8.

3.3 Learning Evolving LFPN in Cascade

In LFPM all behaviors are assumed to be independent with each other, and the correlation among the behaviors is ignored, which will lead to the loss of knowledge carried by the temporal behavior sequence. To minimizing this loss, we can split the evolving social network $\mathbb{G}$ into fine-grained snapshots $G_1, G_2, \cdots, G_t$ and infer the LFPM model on each snapshot, and thus we can build the corresponding LFPN $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_t$ respectively. However, each $\mathcal{G}_t$ is also a snapshot of the evolving $\mathcal{G}$ and only carries the information of the behaviors in t, but ignores the knowledge of the previous behavior history. For learning the complete evolving LFPN, we apply the LFPM in cascade, during which the later models can inherit the knowledge learned from the previous models.

The knowledge accumulation in cascade LFPM is achieved by the prior parameters, which can significantly influence the behaviour of the model by bringing important priori knowledge[10]. Via the priors, the knowledge in the previous period can be transferred in-

Algorithm 2 Building Evolving LFPN

Require:
The evolving network $\mathbb{G}$
Ensure:
The evolving LFPNs $\{\mathcal{G}_t\}$ for $\mathbb{G}$
1: Split $\mathbb{G}$ into T subgraphs by discrete periods
2: **for** each period $t \in T$ **do**
3: Initialize the priors for t using the previous model according to Eq. 9 – 12
4: Infer the LFPM on G_t
5: Initialize LFPN $\mathcal{G}_t$ using the structure of $\mathcal{G}_{t-1}$
6: {In the following we omit the subscript t of weights and distributions for brevity}
7: **for** each ego $u \in G_t$ **do**
8: Add new friends of u in t to $\mathcal{G}_t^{L(u)}$
9: **for** each friend $z \in \mathcal{G}_t^{L(u)}$ **do**
10: Add cross-layer edge $\langle u, z \rangle$ and set $w^E_{u,z} \leftarrow \theta_{u,z}$
11: **end for**
12: **for** each LFP Triple (u, z, v) initiated by u **do**
13: Add local LFP edge $\langle z, v \rangle$ to $\mathcal{G}_t^{L(u)}$
14: Set $w^{L(u)}_{z,v} \leftarrow p(z|\langle u, v \rangle)$
15: **end for**
16: **end for**
17: Add new individuals to $\mathcal{G}_t^G$
18: **for** each LFP Pattern (z, v) derived from the LFP Triples **do**
19: Add global LFP edge $\langle z, v \rangle$ to $\mathcal{G}_t^G$
20: Set $w^G_{z,v} \leftarrow \phi_{z,v}$
21: **end for**
22: **end for**
23: **return** $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_t$

to the next period successively. We define the priors of the LFPM at time t as the linear combination of the accumulated knowledge from previous models encoded by the priors and the new knowledge encoded by the statistics in the last period $t-1$ as follows:

$$\alpha^t_{u,z} = \lambda \cdot \alpha^{t-1}_{u,z} + \Omega^{(u,z)(t-1)}_{u,z} \tag{9}$$

$$\beta^t_{u,y} = \lambda \cdot \beta^{t-1}_{u,y} + \Omega^{(u,v)(t-1)}_{u,y} + \Omega^{(z,y)(t-1)}_{u,y} \tag{10}$$

$$\delta^t_{u,c} = \lambda \cdot \delta^{t-1}_{u,c} + \Omega^{(u,c)(t-1)}_{u,c} \tag{11}$$

$$\gamma^t_{c,u} = \lambda \cdot \gamma^{t-1}_{c,u} + \Omega^{(c,y)(t-1)}_{c,u} \tag{12}$$

where λ is a cascade damping factor. A higher λ means to inherit more priori information from the earlier models. For the first model the priors are initialized with the default values $\alpha_0, \beta_0, \delta_0, \gamma_0$.

While cascade LFPM allows knowledge accumulation in the model inference on the sequential snapshots, the LFPN can also inherit the structures from previous periods. By adding the new structures and updating the weights learned from the LFPM for each period, we can always get the up-to-date evolving LFPN for the evolving social network. The details are given in Algorithm 2.

4. LFPN-BASED LINK PREDICTION

LFPN adopts the transitivity of friendship to explain the evolution of social networks and shows the potential friendship propagation in the network. An interesting problem is, to what extent the LFPN can model the hidden power which drives the growth of social networks. We study this by applying LFPN in the traditional link prediction problem in social networks. We formulate the problem as follows:

LFPN based Link Prediction Problem. Given an LFPN $\mathcal{G}_t$ inferred from a snapshot G_t of the evolving social network G at

Table 2: The statistics of the datasets. V, E: number of existing nodes and edges in each dataset, S: number of egos for prediction, D: total number of candidate friends for the egos, R: total number of candidate friendships for prediction, E/V: average degree of each node, R/S: average candidate friendships of each ego.

Dataset	V	E	S	D	R	E/V	R/S
cond-mat	27,348	72,119	465	21,266	274,817	2.637	1,182.009
hep-ex	5,667	60,425	45	3,476	85,447	10.663	3,797.644
hep-lat	3,804	13,331	216	3,059	74,189	3.504	686.935
hep-ph	17,615	78,932	601	14,943	529,623	4.481	1,762.473
hep-th	14,751	27,299	113	11,608	100,717	1.851	1,782.603
nucl-ex	4,425	59,769	167	3,396	132,803	13.507	1,590.455

time t, our problem is to predict the new friendships in period $t+1$, i.e. ΔE_{t+1}.

Here we don't consider the new-coming vertices in the network. Random Walk (RW) provides an effective node proximity measure for predicting new links [31][1]. However, traditional RW regards the social network as a plain graph while ignoring the motivation of individuals. Unlike them, we predict new links based on behavior modeling. We design a new random walk upon the LFPN, **LFPN-RW**. In LFPN-RW, we treat the creation of new friendship as the result of friendship propagation in the network. In this way, LFPN-RW captures the co-influence effect of the friend circles on one's friend-making behaviors, rather than considering them as independent events. Furthermore, the personal interests are also taken into consideration by being incorporated within the weights of edges.

To predict the new friends of u, consider a random walker who starts the LFPN-RW from the ego u. Then LFPN-RW operates on the *local* and *global* layer of the LFPN successively, modeling the 2-step friend-making process. The walk on *local layer* allows one to find an intermediary first, and to make new friends via this intermediary on the *global layer* next. The basic assumption for the random walk on the *local layer* is that one should prefer the intermediary whom she knows via her preferred intermediaries; and that for the random walk on the *global layer* is that one should be likely to make friends with those recommended by important people.

We introduce the process of LFPN-RW for u as follows. Firstly, the walker goes to any vertex $z \in \mathcal{G}^{L(u)}$ with the intermediary preference probability $w^{E}_{u,z}$, and walks in $\mathcal{G}^{L(u)}$ randomly. At each vertex, the walker can go to next vertex along any local LFP edge with probability $1-\mu$, or go back to the ego u with probability μ and then restart the walk.

When the walker stops at some vertex $z \in \mathcal{G}^{L(u)}$, before jumping to the ego u, she jumps to the projection of z in the global layer $\mathcal{G}^{G}$, and starts the random walk on $\mathcal{G}^{G}$ to find a new friend. Arriving at any vertex in $\mathcal{G}^{G}$, the walker may go to any of its neighbors along the global LFP edges with probability $1-\mu$, or just stop there with probability μ and then jump back to the ego u directly.

Before proceeding, we revise the weight definition of the edges in LFPN to combine the global propagation probabilities and personal interests in the random walk from u:

$$w^{L(u)}_{z,v} = \phi_{z,v} \cdot p(z|\langle u, v\rangle), \tag{13}$$

$$w^{G(u)}_{v,y} = \phi_{v,y} \cdot \sum_{c \in C} (\chi_{u,c} \cdot \psi_{c,y}), \tag{14}$$

where C is the collection of interest areas. Eq. 13 biases the local layer using global propagation probabilities. Meanwhile, Eq. 14 biases the global layer by incorporating the personal interests, and thus ensures that the friendship is more likely to propagate to those with higher reputation in u's preferred interest areas in the random walk of u.

Let $r^{L(u)}(z)$ and $r^{G(u)}(v)$ denote the steady-state probabilities of the random walk initialized from the ego u in $\mathcal{G}^{L(u)}$ and $\mathcal{G}^{G(u)}$, respectively. They satisfy: (assuming all the weights in LFPN are normalized)

$$r^{L(u)}(z) = (1-\mu) \cdot \sum_{z' \in F(u,z)} w^{L(u)}_{z',z} \cdot r^{L(u)}(z') + \mu \cdot w^{E(u)}_{u,z} \tag{15}$$

$$r^{G(u)}(v) = (1-\mu) \cdot \sum_{v' \in F(v)} w^{G(u)}_{v',v} \cdot r^{G(u)}(v') + \mu \cdot r^{L(u)}(v) \tag{16}$$

where μ is the restart probability in the walk. $r^{L(u)}(z)$ can be regarded as the probability that u will select z as the intermediary in a friend-making behavior, and $r^{G(u)}(v)$ as the probability that u decides to make friends with v finally. This can be used to predict or recommend new friends for u.

5. EXPERIMENTAL EVALUATION

In this section, we demonstrate the effectiveness of our model through comprehensive experiments on multiple real-world datasets. We introduce our experimental settings in the first two subsections, and afterwards present and discuss the details of the experimental results.

5.1 Datasets

We construct 6 real-world datasets from Arxiv[1], an archive for electronic preprints of scientific papers in the fields of mathematics, physics, etc. The datasets include **cond-mat** (Condensed Matter), **hep-ex** (High Energy Physics - Experiment), **hep-lat** (High Energy Physics - Lattice), **hep-ph** (High Energy Physics - Phenomenology), **hep-th** (High Energy Physics - Theory) and **nucl-ex** (Nuclear Experiment). For each dataset, we use the network evolution history in 1998–2000 as the training set and predict the new links emerging in 2001–2003. Here we only make predictions for the egos who have been in the network up to 1998. And then for each ego, we select her/his 2-hop neighbors as candidate friends for prediction, because it's found that most of the new links connect two individuals with common friends[16][1]. The statistics of the datasets are shown in Table 2.

5.2 Baselines and Evaluation Methodology

Baselines. We compare our method with other 9 popular link prediction methods, which can be classified into 5 sorts.

- Local neighbourhood based measures, including *Jaccard Coefficient* and *Adamic / Adar*, two best measures reported in [19].
- Variants of random walk based on global network structures, including *Random Walk with Restarts (RWR)* [25] , *SimRank* [13] and *Maximum Entropy Random Walk (MERW)* [4]. *RWR* measures the proximity between two nodes by the probability that one arrives at another in a random walk. *SimRank*

[1]Arxiv: http://arxiv.org/

Table 3: The performance of various link prediction methods on the six datasets.

Dataset	Metric	AdamicAdar	Jaccard	LaFT-Proximity	SimRank	MF	MF-BL	RWR	MERW	SRW	LFPN-RW
cond-mat	MAP	0.20119	0.1743	0.20796	0.12685	0.09929	0.09266	0.25933	0.2165	**0.26378**	0.24867
	P@10	0.08782	0.07815	0.07438	0.05363	0.04761	0.04524	0.1089	0.10901	0.10428	**0.11191**
	AUC	0.69089	0.62261	0.70027	0.5658	0.53794	0.50567	0.74899	0.59948	0.75552	**0.76101**
hep-ex	MAP	0.23527	0.23922	0.23388	0.20722	0.10362	0.12034	**0.26752**	0.20549	0.21131	0.26608
	P@10	0.10001	0.11556	0.07667	**0.12222**	0.06444	0.06667	0.08667	0.05111	0.08889	0.09778
	AUC	0.65977	0.6738	0.67723	0.66678	0.47561	0.47822	0.6697	0.64236	0.66543	**0.67919**
hep-lat	MAP	0.19878	0.15381	0.20324	0.10139	0.06187	0.08202	0.25117	0.19337	0.25893	**0.27711**
	P@10	0.11574	0.08287	0.09083	0.05046	0.02685	0.04352	0.14444	0.08093	0.09028	**0.16019**
	AUC	0.73988	0.67935	0.74152	0.61405	0.51185	0.48524	0.78936	0.61661	0.80564	**0.83025**
hep-ph	MAP	0.17418	0.16228	0.17627	0.09455	0.07149	0.07835	0.24014	0.1994	0.22646	**0.27212**
	P@10	0.08587	0.07056	0.05775	0.03729	0.04344	0.06837	0.11566	0.09274	0.11141	**0.12830**
	AUC	0.71379	0.66339	0.72125	0.59292	0.53282	0.61038	0.76727	0.6312	0.75767	**0.81268**
hep-th	MAP	0.19384	0.16609	0.11209	0.09864	0.08205	0.11605	**0.25992**	0.21746	0.23688	0.25668
	P@10	0.06903	0.06814	0.07918	0.03717	0.02655	0.03982	0.10088	0.10133	0.09425	**0.11239**
	AUC	0.68021	0.63403	0.69231	0.53306	0.51608	0.51852	0.74719	0.60993	0.72813	**0.78204**
nucl-ex	MAP	0.18049	0.17826	0.22870	0.1858	0.09074	0.09723	0.28525	0.29525	0.25826	**0.31502**
	P@10	0.12216	0.10599	0.12994	0.11018	0.0521	0.06527	0.16826	0.16581	0.18497	**0.22515**
	AUC	0.71581	0.67474	0.75488	0.68996	0.49876	0.53021	0.76803	0.72169	0.76481	**0.79627**

estimates how soon two random walkers will meet each other in the graph. While both assume the graph is unweighted, *MERW* assigns the propagation probabilities proportional to the eigenvector centrality of nodes [17].

- A state-of-the-art supervised link prediction approach, *Supervised Random Walk (SRW)* [1], which performs random walk on a weighted graph with the weights learned as a function of features.
- Matrix Factorization methods, including the basic *MF* [30] which considers the hidden features as the recommendation of intermediaries, and an improved method, denoted as *MF-BL* here, which tries to overcome the imbalance problem [22].
- *LaFT-Proximity*, a proximity measure based on the LaFT-Tree, our previous work which studies the transitivity of friendship in social networks[32].

All of the above methods are based on link structures while *SRW* utilizes more external features.

Evaluation Metrics. We evaluate the prediction performance of the above algorithms using 3 metrics: MAP (Mean Average Precision), P@10 and AUC (Area Under the ROC Curve). Generally, MAP measures how well the algorithm ranks positive instances above negative instances, P@10 measures the precision of the top 10 predictions given by the algorithm, and AUC measures how well the algorithm distinguishes positive instances from negative instances.

Parameter Settings. For LFPM, we set the default priors $\alpha_0 = \beta_0 = 1, \delta_0 = \gamma_0 = 0.1$, and the number of interest areas $K = 10$, as they are shown appropriate in our experiments. For each LFPM, we run Gibbs sampling for 20 iterations at most. The cascade damping factor for LFPM $\lambda = 0.5$.

In *LFPN-RW*, *RWR*, *MERW* and *SRW*, we set the restart probability $\mu = 0.15$, and the number of iterations of random walk is limited to 100. For *SimRank*, we set the decay factor $C = 0.8$ and the maximum number of iterations as 6, as in [13]. For *MF* and *MF-BL*, we set the number of hidden features as 20.

For *SRW*, we select the Wilcoxon-Mann-Whitney (WMW) loss function and logical edge strength function, as reported the best in [1]. We select the two-hop neighbors of each ego as the negative instances for that in our settings only the two-hop neighbors are considered as the candidate friends for prediction. We apply the gradient descent method to solve the optimization problem and the iteration is executed at most 20 times. The features for *SRW* include number of friends of the two individuals, number of their common friends, Jaccard coefficient and Adamic/Adar score.

5.3 Performance Comparison

We present the evaluation results in Table 3. As highlighted in bold, *LFPN-RW* consistently outperforms other methods on most datasets and most metrics.

LaFT-Proximity is a proximity measure based on the LaFT-Tree, another work which tries to explain link formation using the transitivity of friendship. Though *LaFT-Proximity* shows better performance than other local neighbourhood based measures, including *Jaccard* and *Adamic/Adar*, it fails when compared with more complex algorithms which can utilize more global network information. *LaFT-Proximity* is always worse than our *LFPN-RW*, demonstrating the importance and effectiveness of modeling interests in LFPM and capturing the co-influence of friend circles in *LFPN-RW*.

By utilizing the global network topological structure for proximity measure, random walk based methods, including *RWR*, *MERW* and *SRW*, show better performance than other baselines. In most occasions their performance exceeds *Adamic/Adar* and *Jaccard*, the two best local measures reported in [19]. However, the variants of *RWR*, *MERW* and *SRW*, are not always better than *RWR* as expected. *SRW* outperforms the *RWR* on **cond-mat** and **hep-lat** on MAP and AUC, but becomes worse on **hep-ph**. *MERW* fails to improve the performance of *RWR* on most of our datasets. Both *MERW* and *SRW* try to weight the edges in the graph based on the topological features; on the contrary, *LFPN-RW* weights the edges by modeling the behaviors of individuals and capturing the co-influence of friend circles based on the common user interests, and thus achieves better improvement over *RWR* than *MERW* and *SRW*. On **hep-lat**, **hep-ph** and **nucl-ex**, *LFPN-RW* achieves on average 0.04 higher AUC and 0.03 higher MAP as compared to *RWR*. Furthermore, we notice that *LFPN-RW* is not only accurate but also more stable than others. This convinces us that the LFPN can capture the important information about the network growth and people's friend-making behaviors.

It is important to note that, unlike *SRW* which requires feature extraction from both positive and negative instances, *LFPN* estimates the weighted network from only the "positive instances", and no other features are considered.

The performance of *LFPN-RW* far exceeds that of *MF* which tries to capture the effect of intermediary by matrix factorization, and that of its improved version *MF-BL*. While in matrix factorization we can explain the latent feature space as the intermediation effect, these methods cannot utilize the internal relations among individuals' friend-making behaviors and thus the intermediation effect cannot be modelled well.

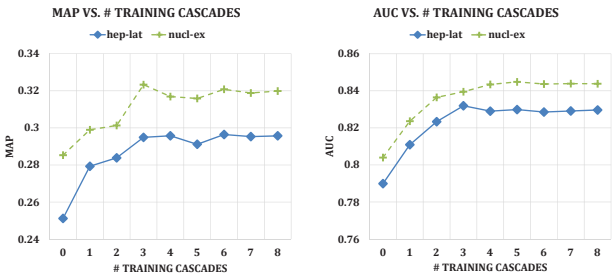
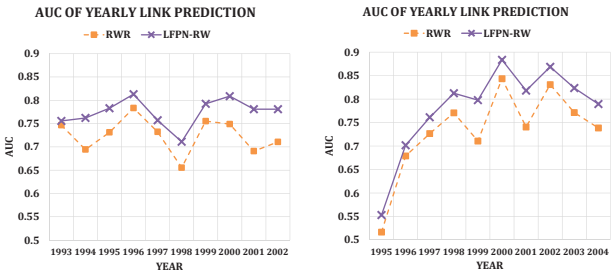


Figure 3: The effect of knowledge accumulation in Cascade LFPM on the link prediction performance of LFPN-RW.



(a) hep-lat (b) nucl-ex

Figure 4: The performance of dynamic link prediction for each year.

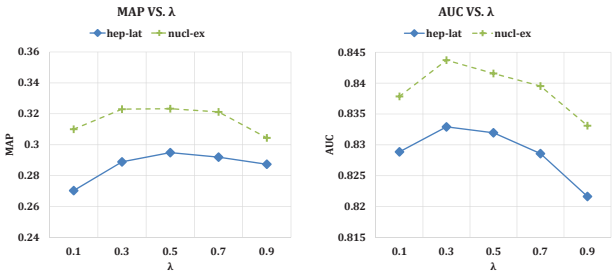


Figure 5: The variance of link prediction performance of LFPN-RW with different damping factor λ in Cascade LFPM.

5.4 Effect of Knowledge Accumulation

In this subsection, we study to what extent the knowledge accumulation in Cascade LFPM can influence the learned LFPN and the link prediction performance based on LFPN. We conduct this study by three questions: (1) Does a longer training time mean more accumulated knowledge and thus promise better performance in link prediction? (2) As time goes by, with more knowledge accumulated in cascades, will the prediction performance increase continuously? (3) How should we control the amount of the accumulated knowledge that be passed from one period to the next?

We answer the first question by examining the performance variance of *LFPN-RW* on the LFPNs with different length of training time. Longer training time means more earlier knowledge is learned and accumulated, and afterwards reflected in the final LFPN. Fig. 3 shows the experimental results. Here we split the training data into multiple cascades by year and each cascade corresponds to one year. For the sake of readability we only show the results on **hep-lat** and **nucl-ex**, and similar results are observed on other datasets. When the number of training cascades is 0, our *LFPN-RW* decays to *RWR*. *LFPN-RW* with one training cascade only utilizes

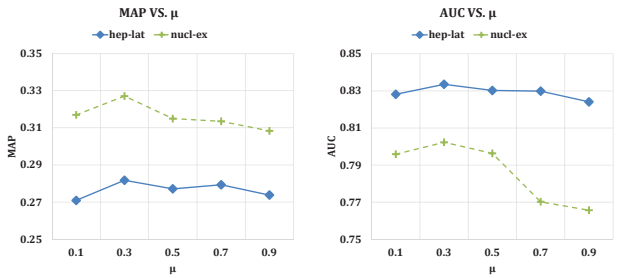


Figure 6: The variance of link prediction performance of LFPN-RW with different μ in LFPN-RW.

the knowledge of the current period and outperforms that with no training cascade, i.e. *RWR*, on both datasets and both metrics. The AUC of *LFPN-RW* increases with the number of training cascades in the beginning, and then tends to be stable after 4 training cascades. This indicates that the accumulated knowledge in the recent time can really improve the performance; however, the knowledge too far in the past has little effect because people's behaviors and interests are changing with time. This finding proves that our approach is more feasible as when only little training data are given it can still show promising performance.

To answer the second question, we conduct dynamic link prediction tasks on the datasets. For fixed individuals in the first year in each dataset, we make yearly link prediction which predicts the new links for the next year given the network in each year, and study how the performance will vary. As time goes by and the number of friends of each individual keeps growing, there will be more candidate friendships than new emerging friendships in each year, so it's difficult to compare the performance on a sequential basis. Thus we compare *LFPN-RW* with *RWR* in each year. The results are presented in Fig. 4. The curves of *RWR* and *LFPN-RW* are quite similar on both datasets. We observe that *LFPN-RW* is not only always better than *RWR*, but also more stable, as shown in the more smooth curve. *RWR* will perform worse if there exist outliers in the network; however, LFPN becomes increasingly robust with continuous knowledge accumulation. Furthermore, in **nucl-ex** and later time of **hep-lat**, we see the improvement of *LFPN-RW* over *RWR* increases with time.

Finally, we study how we can control the extent of knowledge accumulation by the cascade damping factor λ in LFPM inference for better link prediction performance. The choice of λ faces the trade-off between the knowledge learned in the new period and that accumulated from previous models. Just think of the extreme cases. When $\lambda = 0$, each LFPM is trained in each period independently, without the accumulated knowledge; however, when $\lambda = 1$, all previous knowledge is passed to the next model, with equal importance as the new knowledge. Typically, we set a higher λ when less data is observed in the specific period. Otherwise, if we observe sufficient data for the current model, we can lower the influence of the previous models on the current one by setting a smaller λ. When evaluating on real data we observe that λ plays an important role in the cascade inference procedure. For the sake of readability, we only show some representative curves in Fig. 5. We can see λ in the range from 0.3 to 0.5 seems to be most appropriate.

5.5 Effect of Co-influence of Friend Circles

In LFPN-RW, we treat the creation of a new friendship as a result of friendship propagation in the network and model the link prediction as a random walk upon the LFPN, taking the co-influence ef-

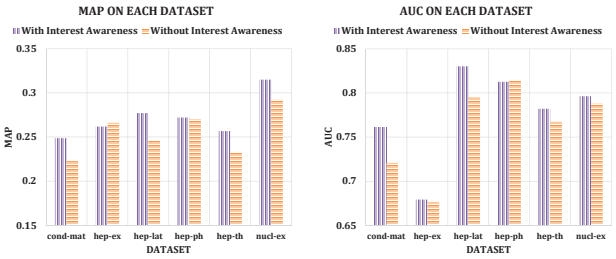


Figure 7: Comparison of link prediction performance of LFPN-RW with/without interest awareness.

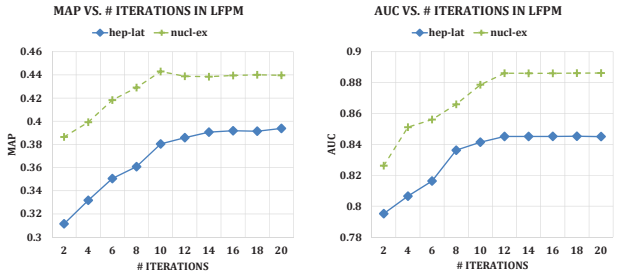


Figure 8: The variance of link prediction performance of LFPN-RW with increasing number of iterations in LFPM.

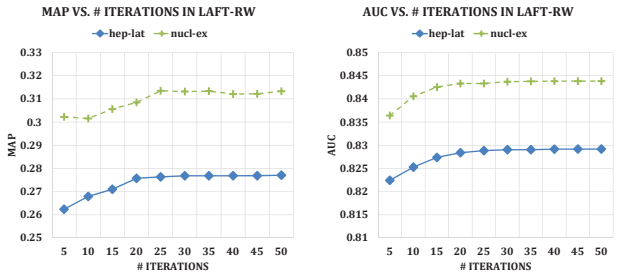


Figure 9: The variance of link prediction performance of LFPN-RW with increasing number of iterations in LFPN-RW.

fect of the friend circles on one's friend-making behaviors into consideration naturally. In this subsection we study how we can benefit from the co-influence effect. In the LFPN-RW, the restart probability μ controls how "far" the walker wanders on the network; a smaller value allows the walker to walk farther and increases the co-influence of the friend circles on one's friend-making behaviors, and a larger value enhances the direct influence of one's current friends. As shown in Fig. 6, in our evaluations we see the lower value of μ tends to achieve better performance. This demonstrates the importance of the co-influence captured in our LFPN-RW. Furthermore, the performance will decrease if $\mu < 0.3$ because the direct influence from the current friends is reduced too much. To achieve better performance, one needs to make trade-off between the influence from current friends and the co-influence of friend circles in specific networks.

5.6 Effect of Interest Awareness

In LFPM and LFPN-RW, we assume that the creation of friendship is the result of friendship propagation, which would be influenced by both the transitivity of friendship and the personal interests. In this subsection, we investigate how personal interests will influence the link prediction performance. We remove the interest factors, including χ and ψ, from LFPM and LFPN-RW, and get a new algorithm denoted by *LFPN-RW without Interest Awareness*. Correspondingly, our original algorithm can be denoted by *LFPN-RW with Interest Awareness*. We compare the performance of the new algorithm with our original algorithm on all the datasets and the results are shown in Fig. 7. We can see the latter outperforms the former on most datasets. It should be noted that we don't introduce external personal profiles to learn the personal interests. This convinces us that the personal interests can be well captured from the social behaviors of the individuals and the topological structure of the network, and furthermore can benefit the link prediction using LFPN-RW, i.e. the *LFPN-RW with Interest Awareness* here, which models the friendship propagation with interest awareness.

5.7 Convergence Analysis

In this subsection we study the convergence speed of our LFPM inference and LFPN-RW.

In Fig. 8 we investigate how the performance of LFPN-RW varies with the number of iterations of LFPM. From the two representative curves on **hep-lat** and **nucl-ex**, we see the performance increases quickly, and then LFPM converges in nearly 15 iterations.

Fig. 9 shows the performance variance with the number of iterations in LFPN-RW. Experimental results on **hep-lat** and **nucl-ex** demonstrate the LFPN-RW converges in about 30 iterations.

6. RELATED WORK

Link prediction is a classical problem which attracts many attentions. The general approach for link prediction is based on the proximity measuring in the network. One branch of this method is based on the local neighbourhood structures, such as *common neighbors*, *Jaccard coefficient* and *Adamic/Adar*, all of which were surveyed by Liben-Nowell and Kleinberg [19]. Another popular approach utilizes the random walk to measure the node proximity on the whole network, including *Random Walk with Restarts (RWR)* [25], *SimRank* [13] and *Katz* [14]. Approaches have also been proposed to improve the traditional random walk by adjusting the transition probabilities , including *randomized shortest-path (RSP) dissimilarity* [29] and *Maximum Entropy Random Walk(MERW)* [4][17]. *Supervised Random Walk(SRW)* was proposed to learn the transition probabilities with a supervised method [1]. However, supervised methods face the imbalance problem [20][22] and require elaborate feature extraction; on the contrary, our LFPN-RW models personal behaviors depending only on the positive instances. Furthermore, we identify personal interests from their social relations only, without other node-specific attributes.

It's also interesting to study how the network evolves [16] and how people make friends [12]. Researchers have found many interesting patterns, such as *preferential attachment* [23], *triadic closure* [27], *reciprocity*[24][11] and *homophily* [5]. The transitivity of friendship has been noticed long ago [26][28] and used to explain the phenomenon of triadic closure in social networks [15]. Recently, the role of triadic closure in the link formation in social networks was further verified [7][27][3][6]. However, it hasn't been studied that how the triadic closure and the transitivity of friendship drive the microscopic evolution of social networks. Yin et al. proposed a matrix factorization approach for link prediction by considering the hidden features as the recommendation of intermediaries [30]. Their work is a bit similar with ours; however, they ignore the sequential relationship among the social behaviors and cannot model the actual contribution of each intermediary on the link formation.

Our previous work [32] primarily focuses on how to capture the expansion traces of one's social network. It does not explore the

dimension of personal interest and model its impact on network expansion. Neither does it utilize the concept of friend circles to capture the co-influence effect of one's friends based on interest awareness as in our paper.

7. CONCLUSION

Modeling people's social behaviors and furthermore understanding their motivations are important for us to know how the social network emerges, evolves and vanishes eventually, as the people are the dominant players in social networks. In this paper, we model people's friend-making behaviors using the LFPM, a generative model driven by the famous sociological principle of the transitivity of friendship and personal interests. The inferred LFPN incorporates rich knowledge about the patterns of individuals's behaviors and the growth potentials of the social network, with the co-influence of friend circles and personal interest well modeled. Furthermore, we propose LFPN-RW, which treats the link prediction task as a random walk on the LFPN, guided by the interest-aware friendship propagation. Our approach achieves promising performance in experimental studies, which leads us to draw the conclusion that the transitivity of friendship really plays important roles in the evolution of social networks and can be utilized to analyze the network evolution if well modeled with interest awareness.

8. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper. This work is supported in part by the National Natural Science Foundation of China (No. 61170064, No. 61073005, No. 61133002), the National High Technology Research and Development Program of China (No. 2012AA011002), US NSF through grants IIS-0905215, CNS-1115234, IIS-0914934, DBI-0960443, and OISE-1129076, and Huawei grant.

9. REFERENCES

[1] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM '11*, pages 635–644, 2011.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[3] M. J. Brzozowski and D. M. Romero. Who should I follow? recommending people in directed social networks. In *ICWSM'11*, 2011.

[4] Z. Burda, J. Duda, J. M. Luck, and B. Waclaw. Localization of the maximal entropy random walk. *Phys Rev Lett*, 102(16):160602, 2009.

[5] M. De Choudhury. Tie formation on twitter: Homophily and structure of egocentric networks. In *SocialCom/PASSAT '11*, pages 465 –470, 2011.

[6] M. Doroud, P. Bhattacharyya, S. F. Wu, and D. Felmlee. The evolution of ego-centric triads: A microscopic approach toward predicting macroscopic network properties. In *SocialCom/PASSAT '11*, pages 172 –179, 2011.

[7] S. A. Golder and S. Yardi. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *SocialCom/PASSAT '10*, pages 88–95, 2010.

[8] M. S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[9] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM '06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

[10] G. Heinrich. Parameter estimation for text analysis. Version 2.9, Fraunhofer IGD, 2009.

[11] J. Hopcroft, T. Lou, and J. Tang. Who will follow you back?: reciprocal relationship prediction. In *CIKM '11*, 2011.

[12] H. Hu and X. Wang. How people make friends in social networking sites - a microscopic perspective. *Physica A: Statistical Mechanics and its Applications*, 391(4):1877 – 1886, 2012.

[13] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *KDD '02*, 2002.

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

[15] D. Krackhardt and M. S. Handcock. Heider vs simmel: emergent features in dynamic structures. In *ICML'06*, 2006.

[16] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08*, 2008.

[17] R.-H. Li, J. X. Yu, and J. Liu. Link prediction: the power of maximal entropy random walk. In *CIKM '11*, 2011.

[18] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03*, 2003.

[19] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7), 2007.

[20] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *KDD '10*, 2010.

[21] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *ANNUAL REVIEW OF SOCIOLOGY*, 27:415–444, 2001.

[22] A. K. Menon and C. Elkan. Link prediction via matrix factorization. In *ECML PKDD'11*, 2011.

[23] M. E. J. Newman. Clustering and preferential attachment in growing networks. *PHYS.REV.E*, 64:025102, 2001.

[24] V.-A. Nguyen, E.-P. Lim, H.-H. Tan, J. Jiang, and A. Sun. Do you trust to get trust? a study of trust reciprocity behaviors and reciprocal trust prediction. In *SDM '10*, 2010.

[25] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *KDD '04*, pages 653–658, 2004.

[26] A. Rapoport. Spread of information through a population with socio-structural bias: I. assumption of transitivity. *Bulletin of Mathematical Biology*, 15:523–533, 1953.

[27] D. M. Romero and J. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *ICWSM '10*, 2010.

[28] G. Simmel and K. H. Wolff. *The Sociology of Georg Simmel*. Free Press, 1964.

[29] L. Yen, M. Saerens, A. Mantrach, and M. Shimbo. A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *KDD '08*, 2008.

[30] D. Yin, L. Hong, and B. D. Davison. Structural link analysis and prediction in microblogs. In *CIKM '11*, 2011.

[31] Z. Yin, M. Gupta, T. Weninger, and J. Han. A unified framework for link recommendation using random walks. In *ASONAM '10*, pages 152–159, 2010.

[32] J. Zhang, C. Wang, J. Wang, and P. S. Yu. LaFT-Tree: Perceiving the expansion trace of one's circle of friends in online social networks. In *WSDM '13*, 2013.