2012 IEEE 12th International Conference on Data Mining

Link Prediction and Recommendation across Heterogeneous Social Networks

Yuxiao Dong*†, Jie Tang*, Sen Wu*, Jilei Tian‡, Nitesh V. Chawla†, Jinghai Rao‡, Huanhuan Cao‡
*Department of Computer Science and Technology, Tsinghua University
†Department of Computer Science and Engineering, University of Notre Dame
‡Nokia Research Center, Beijing
{ydong1,nchawla}@nd.edu, {jietang,senwu}@tsinghua.edu.cn, {jilei.tian,jinghai.rao,happia.cao}@nokia.com

***Abstract*—Link prediction and recommendation is a fundamental problem in social network analysis. The key challenge of link prediction comes from the sparsity of networks due to the strong disproportion of links that they have potential to form to links that do form. Most previous work tries to solve the problem in single network, few research focus on capturing the general principles of link formation across heterogeneous networks.**

In this work, we give a formal definition of link recommendation across heterogeneous networks. Then we propose a ranking factor graph model (RFG) for predicting links in social networks, which effectively improves the predictive performance. Motivated by the intuition that people make friends in different networks with similar principles, we find several social patterns that are general across heterogeneous networks. With the general social patterns, we develop a transfer-based RFG model that combines them with network structure information. This model provides us insight into fundamental principles that drive the link formation and network evolution. Finally, we verify the predictive performance of the presented transfer model on 12 pairs of transfer cases. Our experimental results demonstrate that the transfer of general social patterns indeed help the prediction of links.

***Keywords*-Social network analysis, Link prediction, Recommendation, Factor graph, Heterogeneous networks**

I. Introduction

Social networks are not static. They are dynamic structures that evolve over time either by addition of new vertices or nodes or by new links that form between nodes. Thus, the study and modeling of the dynamics in the network structure are important and a center of focus of a number of papers [1], [2], [3], [4].

In this paper, we consider the process of link formation as a tenet behind network growth and evolution. That is, given nodes in a network, the network grows by forming new relationships among the existing nodes. This has a variety of applications including biology, medicine, and social networks. In this paper, we focus on social networks, considering the question of which individuals will form connections with each other. This is the problem of link prediction or recommendation, which can be defined as the task of predicting whether a link will form between two nodes in the future. However, how such social networks evolve at the level of individual links is still not well understood [5], and forms the main motivation for our work.

Another motivation for this work comes from the major challenge of the link prediction problem which results from the sparsity of real social networks [6], [5], which means that the existing links between nodes are only a very small fraction of all potential links in the network. To solve the strongly unbalanced data between negative instances and positive instances, the authors of [7] undersampled the holdout test set to balance and the authors of [8] also contribute only a sample of the negative instances to their test set. However, this sample changes the data distribution which no longer presents the same challenges at the real-world distribution. This makes the prediction performance is uninterpretable, because it no longer reflects the real capabilities and limitations of the prediction model [6]. [9] studies the problem of inferring the types of social relationships across heterogeneous networks. However, the problem itself is different from the link prediction and recommendation addressed in this work.

While a significant body of research has been conducted on homogeneous social networks, there is very little work on capturing the general principles across heterogeneous social networks. What are the intrinsic mechanisms by which link forms and structure evolves in different social networks? To which extent can we use the general patterns to model the link formation and network evolution? These questions reveal the interacting human behaviors that underlie the fundamental patterns of social activities. The solution to this problem could help shape and improve our understanding of human behaviors and social networks.

The principle of homophily suggests that users with similar characteristics tend to associate with each other [10]. Here we study how four different online networks-Epinions, Slashdot, Wikivote, Twitter-satisfy link homophily, which means that users who share common positive links (trust/friends/vote/reciprocity) will have a tendency to associate with each other. Figure 1 shows the probability of a new relationship exists as a function of the number of common links. Clearly, the likelihood of two users creating a link increases when the number of their common neighbors increases in the four networks. This effect of homophily is more pronounced when the number reaches 100, where the probabilities are all higher than 50% in the four networks. It is worth noting that the probability of reciprocal relations

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.140

in Twitter network increases more sharply than in the other networks in Figure 1.

Let us consider an example in Figure 2. The top part of Figure 2 shows two networks-Twitter and Mobile-which is the input of our problem. The bottom part of Figure 2 is the output of our problem: formation of new links. In Twitter, we try to recommend (or predict) new following links for users and in Mobile network we predict communication relationships. The middle of Figure 2 is the general social patterns we discovered over the two networks for link formation. The fundamental challenge here is how to find the general patterns and bridge them across heterogeneous networks into a model for link prediction.

In this work, we consider the traditional link prediction problem where we split the data into two parts: one is for training and the other for testing. For each user u, we predict users with whom she or he will create a new link and recommend a candidate list for friends in the training data. Then we evaluate whether new links form between u and the recommended candidates in the testing data. Specifically, the paper makes the following contributions.

- We first propose a ranking factor graph model for link prediction, which can extensively improve the performance (on average +10% in terms of AUC and two times higher in $Pre@30$) of friends recommendation over both well-known unsupervised methods and supervised frameworks.
- Then we conduct an investigation of link formation over different online social networks in the high-level of human behaviors. We find some interesting general social patterns in triad relationships, which is the basic unit of network structure across heterogeneous networks.
- Based on the discovered general social patterns, we define the problem of link prediction across heterogeneous networks and propose a transfer-based ranking factor graph model, which incorporates the discovered social patterns into a machine learning framework.

We verify the predictive power of the presented transfer-based model with discovered general social patterns over different networks. Experimental results show that the transfer-based model performs better in most cases when compared to the baseline methods, including our non-transfer ranking factor graph model.

This paper is organized as follows: In Section II, we give a brief description of the online data we use. Section III formulates the problems. Section IV introduces several basic predictors and Section V presents the social patterns and feature definition. Then we propose our transfer-based model and the learning algorithm in Section VI. Section VII introduces the experiment which validates the effectiveness of our model. Finally, we review some previous work related to ours in Section VIII and conclude this work in Section IX.

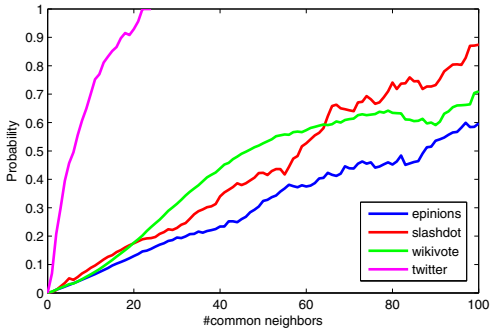


Figure 1. Probability of a new relationship exists as a function of the number of common friends.

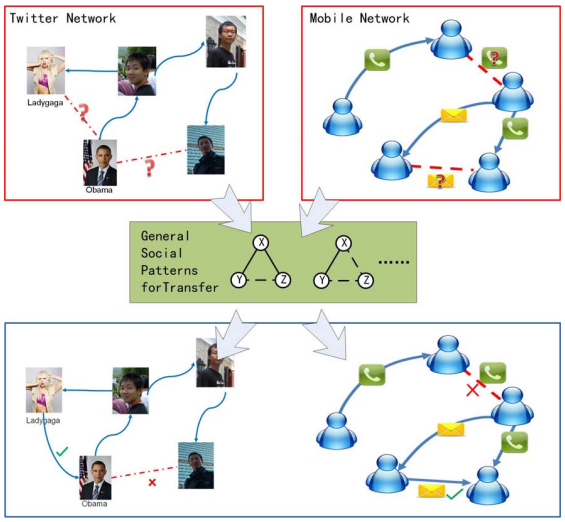


Figure 2. Example of predicting and recommending links across the Twitter network and the Mobile network.

II. Data Description

In this study, we consider four different online social networks from which we can extract positive links: Epinions, Slashdot, Wikivote, Twitter.

Epinions is a who-trust-whom online social web site for product review. The data set consists of 131,828 nodes and 841,372 links, of which about 85.0% are trust links. From this data, we create a network of reviewers with only trust relationships. The trust data set consists of 131,828 nodes and 715,360 links. Our goal on this data set is to predict trust relationships for each user.

Slashdot is a technology-related news website known for its specific user community. The raw network we used here consists of 82,144 nodes and 549,202 links, which was obtained in February 2009 by [11]. From this data, we extract a network with only friendship to recommend friends for each user.

Wikivote is a who-vote-whom network to decide who to promote to adminship in Wikipedia.com. [11] crawled all

Table I
DATASET STATISTICS

Dataset	#nodes	#links	+links	d	cc	di
Epinions	131,828	841,372	85%	13	0.2424	14
Slashdot	82,144	549,202	78%	13	0.0863	11
Wikivote	7,115	103,689	79%	29	0.2089	7
Twitter	63,803	153,098	38%	5	0.1721	24

administrator elections and vote history data. The resulting network contains 7,118 nodes and 103,747 links of which 78.7% are positive. We use the voting links to build our network for positive link recommendation.

Twitter is an online social microblogging network which is built by traversing the following links from 10/12/2010 to 12/23/2010 [4]. On Twitter, when one follows a user, sometimes that user will follow back. Twitter thus facilitates a reciprocal following relationship between users. Here, we use the reciprocal following relationship to build the network for recommending links for each user.

Table I lists statistics of the four networks with positive links we extracted from the original data. d is the average degree of each node and cc is the average clustering coefficient. di denotes the diameter of each network. All the original data is publicly available[1]. Our goal here is to predict and recommend positive relationships for each user.

III. PROBLEM DEFINITION

We first give several necessary definitions, then present the formal definitions of link prediction in singular homogeneous social networks. Finally, we give the problem formulation of link recommendation across heterogeneous networks.

Given a social graph $G(V, E, \mathbf{X})$, where V is a set of $|V| = N$ users and $E \sqsubseteq V \times V$ is a set of friendships among users. $\mathbf{X}$ is a $|V|^2 \times d$ attribute matrix associated with links (both existing links E and non-existing links E^U) in $V \times V$ with each row $\mathbf{x}_s$ corresponding to a link (connecting v_i and v_j), which describes the nodes v_i and v_j (i.e., degree) and the link attributes (i.e., the number of common neighbors, the similarity between v_i and v_j).

Our first goal is to recommend friends for a specific user v_s, based on her/his attributes $\mathbf{x}_s$ and her/his existing friendships. More precisely, we are concerned with the following problem:

Problem 1: **Link Recommendation.** Let $G = (V, E, \mathbf{X})$ be an attribute augmented friendship network. For a particular user v_s and a set of candidates C to which v_s may create a link, the task of recommendation is to find a predictive function such that we can recommend friends for v_s:

$$f : (V, E, \mathbf{X}, v_s, C) \to Y$$

where $Y = \{y_1, y_2, \ldots, y_{|C|}\}$ is a set of inferred results for whether user v_s would create links with users in the candidate set C. The predictive function will output a probability $p(1|e_{si})$ for possible existence between users v_s and v_i; thus our task can be viewed as obtaining a pair $(e_{si}, p(1|e_{si}))$ for each candidate v_i for user v_s.

Then, we turn to the problem of recommendation across heterogeneous networks. The input of this problem consists of two partially labeled networks G_S (source network) and G_T (target network) with $|E_S| >> |E_S^U|$. In other words, there are far more existing links than non-existing links in the source network, with an extreme case of $|E_S^U| = 0$. Based on the traditional link recommendation, we formulate the transfer recommendation with the following format:

Problem 2: **Link Recommendation across Heterogeneous Networks.** Given a source network G_S with abundant positive relationships and a target network G_T, the goal is to learn a predictive function

$$f : (G_T|G_S) \to Y_T$$

for generating the probabilities that a user creates links in the target network by leveraging the information from the source network.

The second problem formulation is different from the traditional link prediction problem [12], [6], [13]. The source and target network could be different. It is also different from the problem of inferring social ties across heterogeneous networks [9], as in this paper we focus on the recommendation problem. What are the fundamental factors that form the link, the micro-clique and the macro-structure of the networks? How reliably can we recommend friends in the target network by using the information provided in the source network? How similar and stable are the behaviors of people when forming friendships in different network worlds?

IV. BASIC PREDICTORS

In this section, we will first describe how we generate candidate relationships and then introduce several baseline predictors.

Candidate Generation. In this work, we try to recommend friends for users. To construct the data, we first randomly selected 2,000 nodes as the source users [5] from the network. For each source user, we generate the candidate list for her/him. Specifically, for a given user v_s, there are in total $(|V| - d(v_s))$ potential links except her/his existing friends.[2] Here, we choose the 2-hop neighborhood as v_s's potential friends. There are two reasons that we focus on predicting links in the distance of 2-hop. First, there is empirically hard to believe that the benefit of successfully predicting links to nodes at high n-hop neighbors is greater than the benefit of predicting them at low n-hop neighbors

[1] Epinions, Slashdot and WikiVote are available at http://snap.stanford.edu. Twitter is available at http://arnetminer.org/reciprocal

[2] $d(v_s)$ denotes the degree of node v_s.

[6]. More than half of all links close triangles at the time of creation, i.e., a person connects to a friend of her/his friend [3]. Second, the number of potential candidates grows exponentially ($d(v_s)^{n-1}$) as the number of hops n increases. For example, in Epinions, if set n=3, then only 0.6% of the candidate relationships will finally be created.

Baseline Predictors. Straightforwardly, we can consider an unsupervised method. In [12], the authors reviewed several typical unsupervised methods for link prediction and found that the *common neighbors (CN)* method, *Adamic/Adar (AA)* [14] measure and *Jaccard (JA) index* have better performance in most cases. Here we consider the three methods as the basic predictors. In addition, we also define another predictor using the *Preferential Attachment (PA)* index. All four methods use the principle of homophily (similarity) to make predictions.

The ***CN*** predictor simply counts the number of common neighbors between two nodes v_i and v_j to make prediction. The ranking score can be formally defined as $|\psi(v_i) \cap \psi(v_j)|$ where $\psi(v)$ denotes the set of neighbors of node v.

The ***AA*** predictor also counts the number of common neighbors, but weights each common neighbor by a measure called rarity, i.e., $\sum_{v_k \in \psi(v_i) \cap \psi(v_j)} \frac{1}{\log d(v_k)}$. Intuitively, if a common neighbor v_k has a large number of neighbors, then it is not a good indicator to connect the candidate node to the given node.

The ***JA*** predictor examines the rate of common neighbors in their neighbors, viz $\frac{\psi(v_i) \cap \psi(v_j)}{\psi(v_i) \cup \psi(v_j)}$.

The ***PA*** predictor calculates the similarity between v_i and v_j by the product of their degrees $d(v_i) \times d(v_j)$.

The unsupervised methods do not use training data, we could further consider the following two supervised methods: SVMRank and logistic regression classification model (LRC), for link recommendation and prediction. SVMRank uses the local attributes associated with each link or node as features to train a classification model and then apply it to rank the potential nodes in the test data. Here, we use SVM-light. LRC uses the same local attributes to train a logistic regression classification model. As for features in the supervised models, we use the same attributes as those on our proposed model (Cf. II).

V. Social Patterns and Feature Definition

In this section, we introduce several interesting social patterns we discovered in the different networks. Based on these patterns, we give the feature definition for supervised methods.

Degree distribution. The power-law distribution indicates that growth and preferential attachment plays an important role in network development [1]. In [15], the authors found the Internet topology fits the power-law relationships. Similarly, we connect the networks used here to power-law distributions. Figure 3 illustrates that all four different online networks can be fitted to power-law distributions, including trust relations in Epinions, friendships in Slashdot, vote relations in WikiVote and reciprocity in the Twitter following network. People with different relations in different networks all prefer to have connections with those who have had more links.

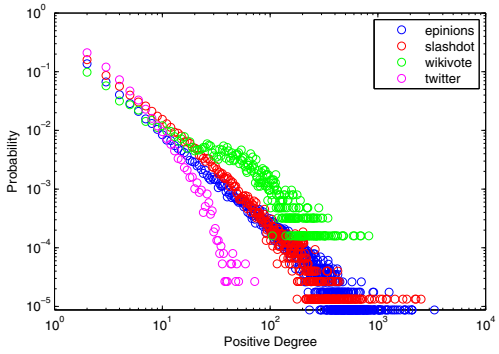


Figure 3. Distribution of the number of positive links (trust / friends / vote / reciprocity) of each user.

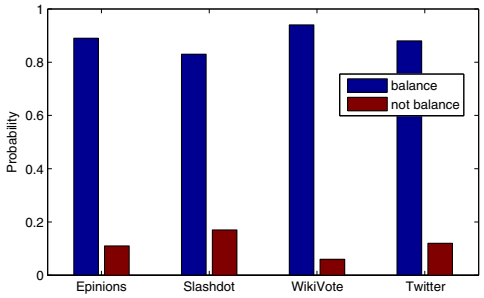


Figure 4. Probabilities of balanced triads in four network based on positive links (trust / friends / vote / reciprocity) and all links.

Social balance. Social balance theory [11] is based on the principles that "the friend of my friend is my friend" and "the enemy of my enemy is my friend", which means that for each group of three users, either all three of these users are friends or only one pair of them are friends. To test whether social balance can help recommendation across heterogeneous networks, we examine how the four different networks satisfy this theory. Figure 4 shows the probabilities of balanced triads based on positive links and all links. It clearly shows that it is more likely (more than 80% likelihood) for users to establish balanced triangle of positive relationships in all four online networks.

Microscopic mechanism. Microsociology is one of the branches of sociology, concerning the nature of human social interactions and agency on a small scale [16]. It shares close association with the philosophy of phenomenology. It can offer us a new perspective to understand the establishment and development of social relationships at a micro-level.

Table II
LOCAL FEATURES DEFINED FOR LINK (v_i, v_j) IN EPINIONS/SLASHDOT/WIKIVOTE/TWITTER.

Feature	description
in-degree	$d_{in}(v_i)$, $d_{in}(v_j)$
out-degree	$d_{out}(v_i)$, $d_{out}(v_j)$
all-degree	$d_{all}(v_i)$, $d_{all}(v_j)$
common neighbors	$\|\psi(v_i) \bigcap \psi(v_j)\|$, $\psi(v_i)$ is the neighbors of v_i
Admic/Adar Index	$\sum_{v_k \in \psi(v_i) \bigcap \psi(v_j)} \frac{1}{\log d(v_k)}$
Jaccard Index	$\frac{\|\psi(v_i) \bigcap \psi(v_j)\|}{\|\psi(v_i) \bigcup \psi(v_j)\|}$
Preferential Index	$d(v_i) \times d(v_j)$

Here, we step from the formation of a close triad, the basic micro-structure in networks to understand the microscopic mechanism of link evolution and network growth. More specifically, we categorize users into two groups (elite users and ordinary users) by estimating the importance of each user by the PageRank algorithm, and selecting the top 1% users [4] as elite users (opinion leaders), with the other as ordinary users. We try to examine the close triad formation with different types of users in it.

Figure 5(a) enumerates six cases of the process of triad formation. We examine the probabilities that two users (Y and Z) have a link, conditioned on whether user X, Y, Z are elite users. There exists some interesting patterns we have found. First, the probabilities of each of the six cases forming a close triad are very distinct. In Figure 5(b), take Epinions network as an example. Conditioned on whether Y and Z are elite users, the probability of Y and Z have a link in case (A/B) (both elite users) is higher (2 to 5 times) than that in (C/D) (either), much higher (10 to 20 times) than that in (E/F) (none). Users Y and Z are more likely (2 to 10 times) to have a link if X is an elite user (A/C/E) than if X is not (B/D/F). Second, the four networks share a very similar distribution on probabilities of close triad formation in all six cases, though the four networks are totally different.

Based on these patterns, we particularly define social pattern-based features (social balance and microscopic mechanism) for the proposed factor graph models. Table II lists a detailed definition of the other features used in the baseline models and the proposed models.

VI. TRANSFER RANKING FACTOR GRAPH MODEL

In this section, we present a (transfer-based) ranking factor graph model (RFG) for friend recommendation.

A. The Link Recommendation Model

To rank the candidates generated for each source user, we propose a ranking factor graph model (RFG) to recommend friends. Figure 6 illustrates the graphical illustration of RFG model.

Given a network $G = (V, E, \mathbf{X})$, the left-bottom figure shows the 2-hop personalized network of a given user v_s and the right figure shows the proposed RFG model. For the TRFG model, the input includes both a source network and a target network in the left part of Figure 6. In the right figure, the graphical model has two layers of variables and two types of functions.

Now we explain the proposed RFG model in detail. For the given user v_s, there are two existing friends: v_5 and v_6. We feed the model with the candidate list $\{v_1, v_2, v_3, v_4\}$ obtained from candidate initialization. The bottom layer of variables in graphical model are observations, which are a collection of both existing and potential friend pairs $\{(v_s, v_i)\}$. The corresponding latent variable in the upper layer represents whether two users are friends. The model incorporates two different types of information including social correlation and attribute correlation. The corporation can be defined as a joint distribution:

$$p(Y|G) = \prod f(v_s, v_i, y_{si}) g(X_c, Y_c) \tag{1}$$

This joint distribution contains two kinds of factor functions which may influence the formation of links.

- **Attribute correlation factor:** $f(v_s, v_i, y_{si})$. It represents the influence of an attribute of potential link between v_s and v_i.
- **Social correlation factor:** $g(X_c, Y_c)$. It denotes the influence of social relation Y_c.

In principle, the two factors can be instantiated in different ways. In this work, we model them by the Hammersley-Clifford theorem [17] in a Markov random field. For the attribute factor, we accumulate all of the attributes and obtain a local entropy for all users:

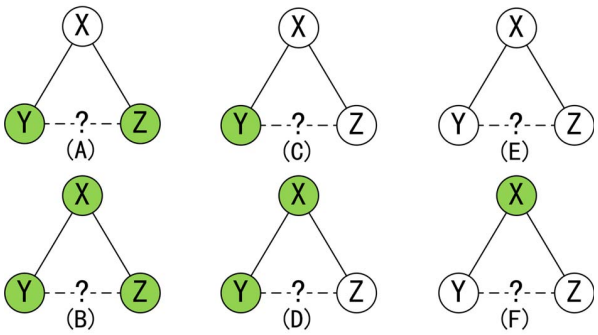
$$f(v_s, v_i, y_{si}) = \frac{1}{Z_\alpha} \exp\{\sum_{j=1}^{d} \alpha_j f_j(x_{si_j}, y_{si})\} \tag{2}$$

where α is the weight of function f_j and Z_α is a normalization factor. It can be defined as either a binary function or a real-value function. For example, for the *common neighbors* feature of node v_s and v_i, we simply define it as a real-value feature.

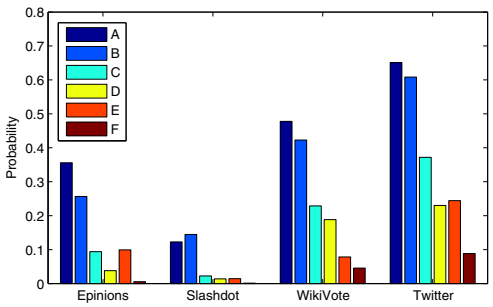
For the social correlation factor, we define a set of correlation feature functions $g_k(X_c, Y_c)$ over each triad Y_c in the network. Then we define a social correlation factor function as follows:

$$\frac{1}{Z_\beta} \exp\{\sum_c \sum_k \beta_k g_k(X_c, Y_c)\} \tag{3}$$

where β_k is the weight of the function, representing the influence degree of k^{th} factor function on Y_c. We take *opinion leader* feature as an example to explain social correlation factor. It is defined as a binary function: if the triad contains an opinion leader, then the value of a corresponding triad factor function is 1, otherwise the value is 0.



(a) Enumeration of triads.



(b) Distribution of six kinds of triads.

Figure 5. Toy example of microscopic mechanism. In 5(a), The enumeration is conditioned on whether X, Y, Z are opinion leaders (green means it is an opinion leader); In 5(b), the Y-axis presents the probabilities that two users (Y and Z) have a link, conditioned on whether user X, Y, Z are opinion leaders.

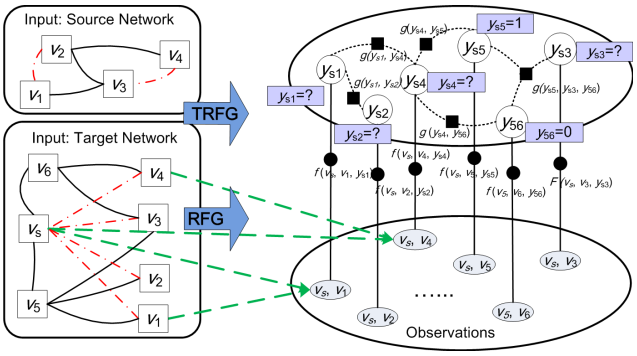


Figure 6. Graphical representation of the RFG and TRFG model. v_s is the given user in the source network who intends to create links with others; $\{v_1, v_2, v_3, v_4\}$ are four candidate friends; v_5 and v_6 are two existing friends of v_s; $\{y_{s1}, \cdots, y_{56}\}$ are latent variables defined for pairs of users, each representing whether the corresponding pair of users will form a friendship; $f(.)$ represents a factor function defined for each pair of users; $g(.)$ represents a correlation factor function defined between latent variables.

where $Z = Z_\alpha Z_\beta$ is a normalization factor. $|V|$ is the set of users to whom we try to recommend friends and $|C|$ is the candidate list for each user.

Finally, by plugging Eqs. 2 and 3 into 1, we define the following log-likelihood objective function $\mathcal{O}(\theta) = \log p(Y|G)$:

$$\mathcal{O}(\theta) = \sum_{s=1}^{|V|}\sum_{i=1}^{|C|}\sum_{j=1}^{d}\alpha_j f_j(x_{si_j}, y_{si}) + \sum_{c}\sum_{k}\beta_k g_k(X_c, Y_c) - \log Z \quad (4)$$

where $Z = Z_\alpha Z_\beta$ is a normalization factor; $|V|$ is the set of users to whom we try to recommend friends and $|C|$ is the candidate list for each user; $\theta = (\{\alpha\}, \{\beta\})$ indicates a parameter configuration.

Learning RFG is to estimate the remaining free parameters θ, which maximizes the log-likelihood objective function $\mathcal{O}(\theta)$. We use the gradient decent method to optimize the objective function. Here we use α as the example to explain how we learn the parameters. Specifically, we first write the gradient of each α_j with regard to the objective $\mathbb{E}$ function,

$$\frac{\partial \mathcal{O}(\theta)}{\partial \alpha_j} = \mathbb{E}[f_j(x_{si_j}, y_{si})] - \mathbb{E}_{P_{\alpha_j}(y_{si}|x_{si_j})}[f_j(x_{si_j}, y_{si})] \quad (5)$$

where $\mathbb{E}[f_j(x_{si_j}, y_{si})]$ is the expectation of feature function $f_j(x_{si_j}, y_{si})$ given the data distribution and $\mathbb{E}_{P_{\alpha_j}(y_{si}|x_{si_j},G)}[f_j(x_{si_j}, y_{si})]$ is the expectation of feature function $f_j(x_{si_j}, y_{si})$ under the distribution $P_{\alpha_j}(y_{si}|x_{si_j})$ given by the estimated model. Usually, it is intractable to estimate the marginal probability in the second term of Eq. 5 as the graphical structure can be arbitrary and may contain cycles. In this work, we use loopy belief propagation (LBP) [18] to approximate the gradients.

B. TRFG Learning

The Transfer-based factor graph model (TranFG) was first proposed by [9]. However, TranFG was designed only for dealing with the classification problem. We now discuss a variant for our recommendation (ranking) problem. Our intuition is that people make friends in different social networks with similar principles. More fundamentally, the formation and evolution of social links which is driven by human behaviors should be general over all social networks. Back to the model, we use the general patterns (social balance and microscopic mechanism) found among different networks and transfer the correlated patterns to help recommend new friends across heterogeneous networks.

We now turn to discuss how to learn the predictive model with two heterogeneous networks (a source network G_S and a target network G_T). Straightforwardly, we can define two separate objective functions for source and target networks. The challenge is then how to bridge the two networks such

that we can transfer the labeled information from the source network to the target network. Therefore, we define the following log-likelihood objective function over the source and target networks by leveraging general patterns of link formation into the proposed TRFG model.

$$\begin{aligned}\mathcal{O}(\alpha,\beta,\mu) &= \mathcal{O}_S(\alpha,\beta) + \mathcal{O}_T(\mu,\beta) \\ &= \sum_{s=1}^{|V_S|}\sum_{i=1}^{|C_S|}\sum_{j=1}^{d}\alpha_j f_j(x^S_{sij}, y^S_{si}) + \sum_{s=1}^{|V_T|}\sum_{i=1}^{|C_T|}\sum_{j=1}^{d'}\mu_j f'_j(x^T_{sij}, y^T_{si}) \\ &\quad + \sum_k \beta_k\Big(\sum_{c\in G_S} g_k(X^S_c, Y^S_c) + \sum_{c\in G_T} g_k(X^T_c, Y^T_c)\Big) - \log Z\end{aligned} \tag{6}$$

where d and d' are the number of attributes in the source network and the target network respectively. In this objective function, the first term and the second term respectively define the likelihood over the source network and the target network; the third term defines the likelihood over common features about social patterns defined in the two networks. The common feature functions are defined according to the general social patterns. Such a definition implies that attributes of the two networks can be entirely different as they are optimized with different parameters $\{\alpha\}$ and $\{\mu\}$, while the information transferred from the source network to the target network is the importance of common features that are defined according to the formation of close triads.

The last issue is to learn the TRFG model. Learning the TRFG model is to estimate a parameter configuration $\theta = (\{\alpha\},\{\beta\},\{\mu\})$ to maximize the log-likelihood objective function $\mathcal{O}(\alpha,\beta,\mu)$. We could still use the gradient decent method to solve the objective function. Detailed learning algorithms for TRFG can be found in [9].

Recommendation. With the estimated parameter θ, the link recommendation is to find the most likely configuration of Y_s for a given user v_s. This can be obtained by:

$$Y^* = \arg\max \mathcal{O}(Y|G,\mathbf{X},\theta) \tag{7}$$

For inference, we use the loopy belief propagation algorithm to find the values of Y_s that maximizes the likelihood. Finally, we can rank the candidate list and recommend friends for the given user.

VII. Experiment

In this section, we first describe evaluation metrics, then present the performance of several baselines and our RFG and TRFG models for recommendation. In the case of TRFG, we finally give several analysis and discussions.

A. Evaluation Metrics

In this work, two-fold cross-validation (i.e., half training and half testing) is used to evaluate the performance of the recommendation and prediction. We quantitatively evaluate the performance of friend recommendation in terms of Precision at Top 30 (*Pre@30*) and the Area under the ROC curve (*AUC*). *AUC* is a related scalar measure of the performance over all thresholds and has classically been used as a measure of performance in link prediction. For *Pre@30*, we evaluate how many of top 30 nodes suggested by the methods actually have links from node v_s. This metric has been used in online social media (i.e., Facebook, Twitter, etc.), where users are presented with a set of friend suggestion.

Our models are implemented in C++, and all experiments were performed on a server running Windows Server 2008 with Intel(R) Xeon CUP E7520 @1.87GHz (16 cores) and 128GB memory. The efficiency performance of the proposed models is acceptable. It takes about three to twenty minutes to train and predict in most cases. For the special case from Epinions to Slashdot network, it takes about two hours, due to the relatively large average degree of each node in these two networks.

B. Experiments without Transfer

First, we illustrate how our RFG can serve as a powerful model for recommending positive links on four different online networks. Table III and IV contain AUC and $Pre@30$ values describing the performance of our model and the other models for predicting potential links within a 2-hop span, namely: *Common Neighbors, Adamic/Adar, Jaccard Index, Preferential Index, SVMRank and Linear Regression Classification Model*. In general, we note that the supervised methods achieve better prediction results than unsupervised ones in terms of both AUC and $Pre@30$, and basically, the AUC values have positive relevance with $Pre@30$. In Table III and IV, it can be clearly seen that our proposed RFG model significantly outperforms both supervised and unsupervised comparison methods. In terms of AUC, RFG achieves a 10-30% improvement compared with the unsupervised methods and SVMRank. Especially in Slashdot, the performance of RFG reaches about 2 times of unsupervised methods. Comparing with LRC method, RFG also gets an improvement of 4-9%. For precision at top 30, there is a slight improvement by RFG to other methods. RFG achieves a better performance than other methods, about 200% relative improvement compared with unsupervised methods in Epinions and Wikivote. In Slashdot, there is a more surprising good performance by RFG. More than 10 out of 30 positive relations we recommend are right in cases of Slashdot and Wikivote. One of the reasons that our RFG has better performance is that it considers some implicit social patterns, namely social balance and close triad formation.

C. Experiments across Heterogeneous Networks

Performance analysis. We now describe the transfer power of link recommendation on 12 pairs of networks: Slashdot/Wikivote/Twitter as source network to Epinions as

target network, Epinions/Wikivote/Twitter (S) to Slashdot (T), Epinions/Slashdot/Twitter (S) to Wikivote (T), Epinions /Slashdot/Wikivote (S) to Twitter (T). In all experiments, the number of positive instances in the source network is set to 50% n (n is the number of negative instances in target network) and the number of source negative instances is equal to 50% target positive instances.

Table V shows the recommendation results on 12 transfer cases. We can see that TRFG with information transfer from source networks outperforms the RFG without transfer in most cases. With Epinions, the transfer from all three other networks: Slashdot, Wikivote, Twitter benefits the recommendation performance evaluated by both AUC and $Pre@30$ metrics. With Slashdot, only the transfer from Epinions improves the predictive power. We also note that all the transfers from Epinions or Slashdot have a more powerful prediction than RFG. With Twitter, TRFG has a slight advantage over RFG.

Although TRFG outperforms RFG in most cases by leveraging the supervised information from source networks, it is worth noting that the transfer seems unhelpful especially for the cases from Twitter or Wikivote network. By a careful consideration, we notice that the degree distributions of Twitter and Wikivote networks does not fully fit the degree distribution like Epinions and Slashdot in Figure 3. There indeed exist some similar patterns over different networks, but the generality of them is sometimes limited.

Transfer efficiency analysis. Here we try to test how the scale of source networks can affect the prediction performance. Figure 7 shows the performance of link recommendation with transfer by varying the percent of source instances to target instances. In general, it shows that the prediction performance decreases as the percent increases, which is contrary to our intuition that more source instances could help solve the imbalance problem in target data. The reason for it is interpretable. Although the social patterns across heterogeneous networks are general, the original distribution of the target network varies to some certain extent with more and more source instances, which may indeed help decrease the imbalance ratio between negative and positive instances. Therefore, to what extent we can leverage the information transferred from source networks depends on the specific pairs of networks. For example, with increasing instances in source networks, the performance from Slashdot to Epinions improves. In contrast, performance decreases when transfer from Wikivote or Twitter networks.

Table III
PREDICTION WITHOUT TRANSFER, EVALUATED BY AUC

Method	Epinions	Slashdot	Wikivote	Twitter
CN	0.8728	0.5048	0.7842	0.5920
AA	0.8736	0.5362	0.7924	0.6198
JA	0.6850	0.3277	0.7241	0.5718
PA	0.8300	0.7108	0.7433	0.5725
SVMRank	0.8943	0.7880	0.7907	0.6834
LRC	0.9405	0.9200	0.8905	0.8044
RFG	**0.9821**	**0.9866**	**0.9298**	**0.8905**

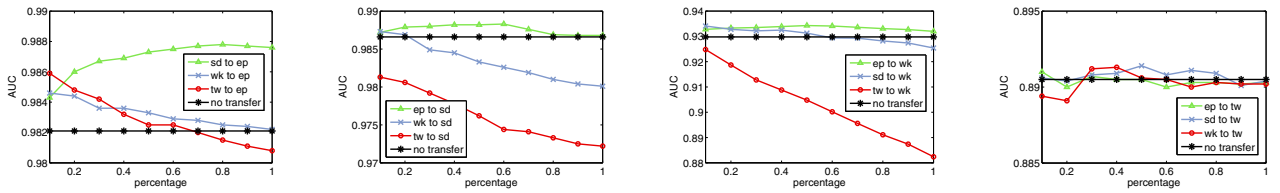
Table IV
PREDICTION WITHOUT TRANSFER, EVALUATED BY PRECISION@30

Method	Epinions	Slashdot	Wikivote	Twitter
CN	3.00	1.81	5.49	4.53
AA	3.38	2.29	5.76	4.64
JA	2.62	1.10	3.45	4.61
PA	1.95	1.76	3.29	4.79
SVMRank	3.14	4.33	7.95	3.37
LRC	4.62	8.14	10.08	4.54
RFG	**5.38**	**11.96**	**12.02**	**5.07**

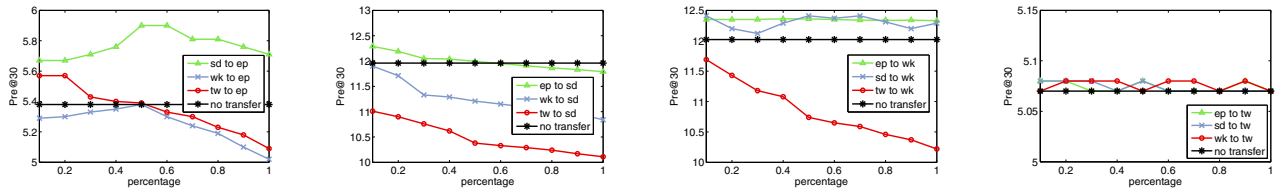
VIII. RELATED WORK

In this section, we review related work on link prediction, social behavior analysis and transfer learning.

Link Prediction. Link prediction has attracted considerable attention in recent years from both the computer science and physics community. Existing work can be classified into two categories: unsupervised methods and supervised methods. Most unsupervised link prediction algorithms are based on the similarity measure between nodes of a graph. A seminal work by Liben-Nowell and Kleinberg for unsupervised methods addresses the problem from an algorithmic point of view, investigating how different proximity features can be exploited to predict the occurrence of new links in social networks [12]. More recently, researchers have advocated supervised approaches for link prediction. [19] proposes a partially labeled factor graph for learning to predict the type of social relationships in large networks. [9] further extends this work to heterogeneous networks by leveraging social theories as the bridge to connect different networks. In [20], Wang et. al. introduce a local probabilistic graphical model that can scale to large graphs to estimate the joint co-occurrence probability of two nodes. [21] presents a unified framework for learning link prediction and edge weight prediction functions in large networks, based on the transformation of a graph's algebraic spectrum. In [6], [22], Lichtenwalter et al. motivate the use of a binary classification framework and vertex collocation profiles through a careful investigation of many factors. In [5], a supervised random walk is designed for link prediction and recommendation in Facebook. The main difference between traditional work on link prediction and our direction lies in that existing work mainly focuses on specific networks, while we try to exploit the general social patterns across heterogeneous networks and incorporate them into a transfer-based ranking factor graph model.



(a) Source to Epinions by AUC (b) Source to Slashdot by AUC (c) Source to Wikivote by AUC (d) Source to Twitter by AUC



(e) Source to Epinions by Pre@30 (f) Source to Slashdot by Pre@30 (g) Source to Wikivote by Pre@30 (h) Source to Twitter by Pre@30

Figure 7. Performance of link recommendation with transfer by varying the percent of source instances to target instances. (ep means epinions network, sd means slashdot network, wk means wikivote network and tw means twitter network.)

Table V
PREDICTION PERFORMANCE OF DIFFERENT TRANSFER CASES EVALUATED BY AUC AND PRE@30. (S) INDICATES THE SOURCE NETWORK AND (T) THE TARGET NETWORK.

Transfer cases	AUC	Pre@30
Epinions	0.9821	5.38
Slashdot (S) to Epinions (T)	**0.9873**	**5.90**
Wikivote (S) to Epinions (T)	**0.9833**	5.38
Twitter (S) to Epinions (T)	**0.9825**	**5.39**
Slashdot	0.9866	11.96
Epinions (S) to Slashdot (T)	**0.9882**	**11.99**
Wikivote (S) to Slashdot (T)	0.9833	11.21
Twitter (S) to Slashdot (T)	0.9762	10.38
Wikivote	0.9298	12.02
Epinions (S) to Wikivote (T)	**0.9343**	**12.36**
Slashdot (S) to Wikivote (T)	**0.9313**	**12.41**
Twitter (S) to Wikivote (T)	0.9048	10.74
Twitter	0.8905	5.07
Epinions (S) to Twitter (T)	0.8905	**5.08**
Slashdot (S) to Twitter (T)	**0.8914**	**5.08**
Wikivote (S) to Twitter (T)	**0.8906**	5.07

Social Behavior Analysis. Our work is also related with social behavior analysis, because we try employ the general human behaviors in online social networks. Barabási et al. [23], [24], [13] take a lot of work to understand individual human behavior patterns, model the scaling properties of human sociality and study structure, social ties and predictability in communication networks. Eagle et al. [25] have considered how interactions between people over mobile communication can accurately predict relations among them. The authors of [26] investigate how social actions evolve in a dynamic social network and propose a time-varying factor graph model for modeling and predicting users' social behaviors. Hopcroft et al. [4] investigate how social theory influences the formation of Twitter network in high levels. Tang et al. [27] study how collaboration relationships have been formed across different domains. Again, existing work focuses on social behavior analysis in the same network, while our work here tries to connect some general social patterns over different networks.

Transfer Learning. Another type of related work is transfer learning, which aims to transfer knowledge from a source domain to a related target domain. Two main issues in transfer learning are "what to transfer" and "when to transfer" [28]. Many approaches have been proposed by selecting instances from the source domain for reusage in the target domain [29], [30]. There is a lot of work conducted to transfer features between different domains. For example, Argyriou and Evgeniou [31] propose a method to learn a shared low-dimensional representation for multiple related tasks. In recent years, there is some work about transferring knowledge across heterogeneous feature spaces [32]. For example, Argyriou et al. [33] propose an algorithm for classification in a heterogeneous environment. Compared with existing work, the networks studied in our problem are quite different and may not even have any overlapping attribute features, while most existing works only consider homogeneous networks. Also, we combine general social features into a transfer learning framework, while existing methods are mainly concerned with how to find shared attributes across different domains.

IX. CONCLUSION

In this work, we study the problem of link prediction and link recommendation both in homoeonomous networks and across heterogeneous networks. First, we precisely define the problems. Then we propose a ranking factor graph (RFG)

model for traditional link prediction and a transfer-based RFG (TRFG) for the novel link-recommendation problem across heterogeneous networks. At the micro-level, we find several general social patterns over different online social networks. We combine the discovered general social patterns into TRFG, which is used to transfer supervised information from the source network to help predict and recommend links in the target network. Experimental results in both cases show that our presented models can significantly improve the predictive performance by comparing them with several baseline methods.

Although the scale of online social networks is growing at an exponential rate, the microscopic mechanism of link formation is still largely unexplored. Exploring the general social patterns of link formation could help us understand human interactions better. There are many potential future directions of this work. First, other general social patterns can be further explored. Another idea is to apply the proposed methodologies to other social networks to further validate its effectiveness.

Acknowledgements. This work was done when the first author was visiting Tsinghua University. The work is supported by the Natural Science Foundation of China (No. 61222212, 61073073), Chinese National Key Foundation Research (No. 60933013, No.61035004), U.S. National Science Foundation (NSF) Grant BCS-0826958 and the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. It is also supported by a research funding from Nokia Research Center.

REFERENCES

[1] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, 1999.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: membership, growth, and evolution," in *KDD'06*, 2006, pp. 44–54.

[3] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," in *KDD'08*, 2008, pp. 462–470.

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

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

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

[7] S. S. Mohammad AI Hasan, Vineet Chaoji and M. Zaki, "Link prediction using supervised learning," in *Workshop on LACS of SDM'06*, 2006, pp. 322–331.

[8] C. Wang, V. Satuluri, and S. Parthasarathy, "Local probabilistic models for link prediction," in *ICDM'07*, 2007, pp. 322–331.

[9] J. Tang, T. Lou, and J. Kleinberg, "Inferring social ties across heterogenous networks," in *WSDM'12*, 2012, pp. 743–752.

[10] P. F. Lazarsfeld and R. K. Merton, "Friendship as a social process: A substantive and methodological analysis," *Freedom and control in modern society*, pp. 8–66, 1954.

[11] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *WWW'10*, 2010, pp. 641–650.

[12] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *CIKM'03*. ACM, 2003.

[13] D. Wang, Dashun. Pedreschi and A. Barabási, "Human Mobility, Social Ties, and Link Prediction," in *KDD '11*. ACM, 2011.

[14] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *SOCIAL NETWORKS*, vol. 25, pp. 211–230, 2001.

[15] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *SIGCOMM'99*, 1999, pp. 251–262.

[16] N. J. Smelser, *Problematics of Sociology*. The Georg Simmel Lectures, 1995.

[17] J. M. Hammersley and P. Clifford, "Markov field on finite graphs and lattices," *Unpublished manuscript*, 1971.

[18] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *UAI'99*, 1999, pp. 467–475.

[19] W. Tang, H. Zhuang, and J. Tang, "Learning to infer social ties in large networks," in *ECML/PKDD'11*, 2011, pp. 381–397.

[20] C. Wang, V. Satuluri, and S. Parthasarathy, "Local probabilistic models for link prediction," in *ICDM'07*, 2007, pp. 322–331.

[21] J. Kunegis and A. Lommatzsch, "Learning spectral graph transformations for link prediction," in *ICML'09*, 2009.

[22] R. N. Lichtenwalter and N. V. Chawlai, "Vertex collocation profiles: subgraph counting for link analysis and prediction," in *WWW'12*, 2012.

[23] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, 2008.

[24] C. Song, T. Koren, P. Wang, and A.-L. Barabasi, "Modelling the scaling properties of human mobility," *Nature Physics*, 2010.

[25] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring social network structure using mobile phone data," *PNAS*, vol. 106, no. 36, 2009.

[26] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang, "Social action tracking via noise tolerant time-varying factor graphs," in *KDD'10*, 2010, pp. 1049–1058.

[27] J. Tang, S. Wu, J. Sun, and H. Su, "Cross-domain collaboration recommendation," in *KDD'12*, 2012, pp. 1285–1293.

[28] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE TKDE*, 2010.

[29] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th international conference on Machine learning (ICML'07)*, 2007, pp. 193–200.

[30] J. Gao, W. Fan, J. Jian, and J. Han, "Knowledge transfer via multiple model local structure mapping," in *KDD'08*, 2008, pp. 283–291.

[31] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *NIPS'06*, 2006, pp. 41–48.

[32] X. Ling, G. Xue, W. Dai, Y. Jiang, Q. Yang, and Y. Yu, "Can chinese web pages be classified with english data source?" in *WWW'08*, 2008, pp. 969–978.

[33] A. Argyriou, A. Maurer, and M. Pontil, "An algorithm for transfer learning in a heterogeneous environment," in *ECML/PKDD'08*, 2008, pp. 71–85.