Inferring Social Ties across Heterogenous Networks

Jie Tang
Department of Computer Science
Tsinghua University
Beijing 100084, China
jietang@tsinghua.edu.cn

Tiancheng Lou
Institute for Interdisciplinary Information Sciences
Tsinghua University
Beijing 100084, China
ltc08@tsinghua.edu.cn

Jon Kleinberg
Department of Computer Science
Cornell University
Ithaca NY 14853
kleinber@cs.cornell.edu

ABSTRACT

It is well known that different types of social ties have essentially different influence between people. However, users in online social networks rarely categorize their contacts into "family", "colleagues", or "classmates". While a bulk of research has focused on inferring particular types of relationships in a specific social network, few publications systematically study the generalization of the problem of inferring social ties over multiple heterogeneous networks. In this work, we develop a framework for classifying the type of social relationships by learning across heterogeneous networks. The framework incorporates social theories into a machine learning model, which effectively improves the accuracy of inferring the type of social relationships in a target network, by borrowing knowledge from a different source network. Our empirical study on five different genres of networks validates the effectiveness of the proposed framework. For example, by leveraging information from a coauthor network with labeled advisor-advisee relationships, the proposed framework is able to obtain an F1-score of 90% (8-28% improvements over alternative methods) for inferring manager-subordinate relationships in an enterprise email network.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining; H.2.8 [**Database Management**]: Database Applications

General Terms

Algorithms, Experimentation

Keywords

Social network, Predictive model, Social influence

1. INTRODUCTION

Our real social networks are complex and consist of many overlapping parts. Nobody exists merely in one social network. People are connected via different types of social ties in different networks. For example, in an enterprise email network, where people are connected by sending/receiving emails to/from others, the relationships between people can be categorized as manager-subordinate, colleague, etc.; in a mobile communication network, the relationship types could include family, colleagues, and friends. It is well known that the different types of social ties have essentially different influence between people. A graduate's research topic may be mainly influenced by his or her advisor, while other parts of his everyday life will be more influenced by their close friends. Awareness of these different types of social relationships can benefit many applications. For example, if we could have extracted friendships between users from a mobile communication network, we can leverage the friendships for a "word-of-mouth" promotion of a new product.

However, in most online networks (e.g., Facebook, Twitter, LinkedIn, YouTube, and Slashdot), such information (relationship type) is usually unavailable. Users may easily add links to others by clicking "friend request", "follow" or "agree", but do not often take the time to create labels and maintain their friend list. Indeed, one survey of mobile phone users in Europe shows that only 16% of users have created contact groups on their mobile phones [10, 27]; our preliminary statistics on LinkedIn data also shows that more than 70% of the connections have not been well labeled. A few efforts have been made to infer social ties. For example, Crandall et al. [4] investigate the problem of inferring friendships between people from co-occurrence in time and space. Wang et al. [30] aim to discover advisor-advisee relationships from the publication network. Diehl et al. [6] try to identify social ties (e.g., manager-subordinate) by learning a ranking function with predefined features. However, most of these works focus on mining particular types of relationships in a specific domain. For example, [30] defines two heuristic rules as constraints and tries to discover advisor-advisee relationships by propagating the constraints in a graphical model. However, the method is difficult to extend to other domains.

Another challenge is that different networks are very unbalanced. In some networks, such as Slashdot, it might be easy to collect some labeled relationships (e.g., trust/distrust relationships between users). However, in most other networks, it may be infeasible to obtain the labeled information and thus difficult to accurately infer the social relationships. One potential opportunity is that in the real world, different networks are intertwined with, instead of separated from, each other. Can we leverage the correlations between different networks to help infer the types of social ties?

Motivating Examples To clearly illustrate the problem, Figure 1 gives an example of inferring social ties across a product reviewers' network and a mobile communication network. In Figure 1, the left sub-figure is the input to our problem: a reviewer network, which consists of reviewers and relationships between reviewers; and a mobile network, which is comprised of mobile users and their communication information (calling or texting message). The

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...$5.00.

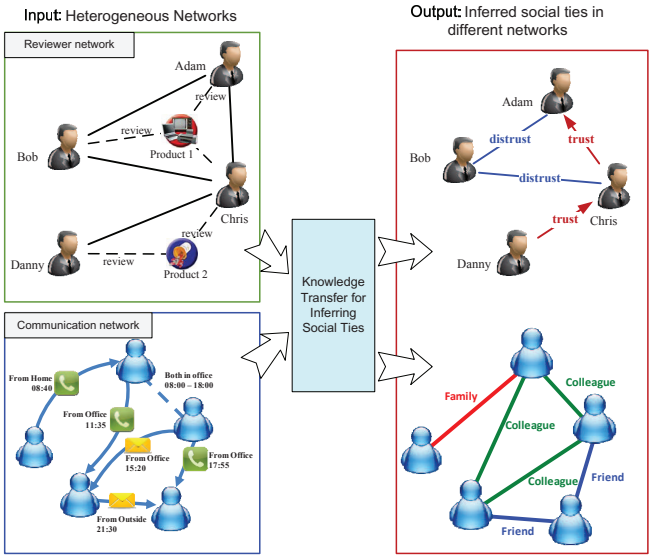


Figure 1: Example of inferring social ties across two heterogeneous networks: a reviewer network and a mobile communication network.

right sub-figure shows the output of our problem: the inferred social ties in the two networks. In the reviewer network, we infer the trust/distrust relationships and in the communication network, we identify friendships, colleagues, and families. The middle of Figure 1 is the component of knowledge transfer for inferring social ties in different networks. This is the key objective of this work. The fundamental challenge is how to bridge the available knowledge from different networks to help infer the different types of social relationships.

The problem is non-trivial and poses a set of unique challenges. First, what are the fundamental factors that form the structure of different networks? Second, how can we design a generalized framework to formalize the problem in a unified way? Third, as real social networks are getting larger with hundreds of millions of nodes, how to scale up the model learning algorithm to adapt to the growth of large real networks?

Results In this work, we aim to conduct a systematic investigation of the problem of inferring social ties across heterogeneous networks. We precisely define the problem and propose a transfer-based factor graph (TranFG) model. The model incorporates social theories into a semi-supervised learning framework, which can be used to transfer supervised information from a source network to help infer social ties in a target network.

We evaluate the proposed model on five different genres of networks: Epinions, Slashdot, Mobile, Coauthor, and Enron. We show that the proposed model can significantly improve the performance (averagely +15% in terms of F1-Measure) for inferring social ties across different networks comparing with several alternative methods. Our study also reveals several interesting phenomena for social science:

- Social balance is satisfied well on friendship (or trust) networks; but not ($< 20\%$ with a large variance) on user communication networks (e.g., mobile communication network).
- Users are more likely (+10%-+98% higher than chance) to have the same type of relationship with a user who spans a structural hole. Disconnected users have an even higher likelihood.
- It was validated that social status is satisfied in many networks. We further discover that several frequent forms of triads have a similar distribution in different networks (Coauthor and Enron).
- Opinion leaders are more likely (+71%-+84%) to have a higher social status than ordinary uses.

Organization Section 2 formulates the problem; Section 3 introduces the data set and our observations over different networks. Section 4 explains the proposed model and describes the algorithm for learning the model; Section 5 gives the experimental setup and Section 6 presents the results; finally, Section 7 discusses related work and Section 8 concludes.

2. PROBLEM DEFINITION

In this section, we first give several necessary definitions and then present the problem formulation. To simplify the explanation, we frame the problem with two social networks: a source network and a target network, although the generalization of this framework to multiple network setting is straightforward.

Let $G = (V, E^L, E^U, \mathbf{X})$ denote a partially labeled social network, where E^L is a set of labeled relationships and E^U is a set of unlabeled relationships with $E^L \cup E^U = E$; $\mathbf{X}$ is an $|E| \times d$ attribute matrix associated with edges in E with each row corresponding to an edge, each column an attribute, and an element x_{ij} denoting the value of the j^{th} attribute of edge e_i. The label of edge e_i is denoted as $y_i \in \mathcal{Y}$, where $\mathcal{Y}$ is the possible space of the labels (e.g., family, colleague, classmate).

Input: The input to our problem consists of two partially labeled networks G_S (source network) and G_T (target network) with $|E_S^L| \gg |E_T^L|$. In other words, the number of labeled relationships in the source network is more larger than that of the target network, with an extreme case of $|E_T^L| = 0$.

In real social networks, the relationship could be undirected (e.g., friendships in a mobile network) or directed (e.g., manager-subordinate relationships in an enterprise email network). To keep things consistent, we will first introduce the problem in the context of undirected network and then discuss how to extend the proposed framework to the directed ones. In addition, the label of a relationship may be static (e.g., the family-member relationship) or change over time (e.g., the manager-subordinate relationship). In this work, we focus on static relationships.

Learning Task: Given a source network G_S with abundantly labeled relationships and a target network G_T with a limited number of labeled relationships, the goal is to learn a predictive function $f : (G_T|G_S) \rightarrow Y_T$ for inferring the type of relationships in the target network by leveraging the supervised information (labeled relationships) from the source network.

Without loss of generality, we assume that for each possible type y_i of relationship e_i, the predictive function will output a probability $p(y_i|e_i)$; thus our task can be viewed as obtaining a triple $(e_i, y_i, p(y_i|e_i))$ to characterize each link e_i in the social network. There are several key issues that make our problem formulation different from existing works on social relationship mining [4, 6, 29, 30]. First, the source network and the target network may be very different, e.g., a coauthor network and an email network. What are the fundamental factors that form the structure of the networks? Second, the label of relationships in the target network and that of the source network could be different. How reliably can we infer the labels of relationships in the target network using the information provided by the source network? Third, as both the source and

the target networks are partially labeled, the learning framework should consider the labeled information as well as the unlabeled information.

3. DATA AND OBSERVATIONS

3.1 Data Collection

We try to find a number of different types of networks to investigate the problem of inferring social ties across heterogenous networks. In this study, we consider five different types of networks: Epinions, Slashdot, Mobile, Coauthor, and Enron. Table 1 lists statistics of the five networks. All data sets and codes used in this work are publicly available.[1]

Epinions is a network of product reviewers. Each user on the site can post a review on any product and other users would rate the review with trust or distrust. In this data, we created a network of reviewers connected with trust and distrust relationships. The data set consists of 131,828 nodes (users) and 841,372 edges, of which about 85.0% are trust links. 80,668 users received at least one trust or distrust edge. Our goal on this data set is to infer the trust relationships between users.

Slashdot is a network of friends. Slashdot is a site for sharing technology related news. In 2002, Slashdot introduced the Slashdot Zoo which allows users to tag each other as "friends" (like) or "foes" (dislike). The data set is comprised of 77,357 users and 516,575 edges of which 76.7% are "friend" relationships. Our goal on this data set is to infer the "friend" relationships between users.

Mobile is a network of mobile users. The data set is from [7]. It consists of the logs of calls, blue-tooth scanning data and cell tower IDs of 107 users during about ten months. If two users communicated (making a call or sending a text message) with each other or co-occurred in the same place, we create an edge between them. In total, the data contains 5,436 edges. Our goal is to infer whether two users have a friend relationship. For evaluation, all users are required to complete an online survey, in which 157 pairs of users are labeled as friends of each other.

Coauthor is a network of authors. The data set, crawled from Arnetminer.org [28], is comprised of 815,946 authors and 2,792,833 coauthor relationships. In this data set, we attempt to infer advisor-advisee relationships between coauthors. For evaluation, we created a smaller ground truth data in the following ways: (1) collecting the advisor-advisee information from the Mathematics Genealogy project[2] and the AI Genealogy project[3]; (2) manually crawling the advisor-advisee information from researchers' homepages. Finally, we have created a data set with 1,534 coauthor relationships, of which 514 are advisor-advisee relationships.

Enron is an email communication network. It consists of 136,329 emails between 151 Enron employees. Two types of relationships, i.e., manager-subordinate and colleague, were annotated between these employees. The data set was provided by [6].Our goal on this data set is to infer manager-subordinate relationships between users. There are in total 3,572 edges, of which 133 are manager-subordinate relationships.

Please note that for the first three data sets (i.e., Epinions, Slashdot, and Mobile), our goal is to infer undirected relationships (friendships or trustful relationships); while for the other two data sets (i.e., Coauthor and Enron), our goal is to infer directed relationships (the source end has a higher social status than the target end, e.g., advisor-advisee relationships and manager-subordinate relationships).

[1]http://arnetminer.org/socialtie/

[2]http://www.genealogy.math.ndsu.nodak.edu

[3]http://aigp.eecs.umich.edu

Table 1: Statistics of five data sets.

Relationship	Dataset	#Nodes	#Edges
Trust	Epinions	131,828	841,372
Friendship	Slashdot	77,357	516,575
Friendship	Mobile	107	5,436
Advisor-advisee	Coauthor	815,946	2,792,833
Manager-subordinate	Enron	151	3,572

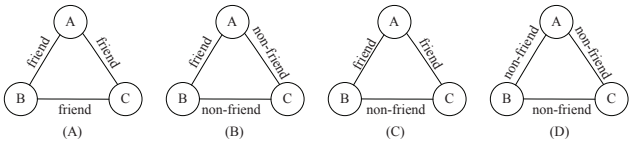


Figure 2: Illustration of structural balance theory. (A) and (B) are balanced, while (C) and (D) are not balanced.

3.2 Observations

As a first step, we engage in some high-level investigation of how different factors influence the formation of different social ties in different networks. Generally, if we consider inferring particular social ties in a specific network (e.g., mining advisor-advisee relationships from the publication network), we can define domain-specific features and learn a predictive model based on some training data. The problem becomes very different, when handling multiple heterogeneous networks, as the defined features in different networks may be significantly different. To solve this problem, we connect our problem to several basic social psychological theories and focus our analysis on the network based correlations via the following statistics:

1. *Social balance [8].* How is the social balance property satisfied and correlated in different networks?
2. *Structural hole [3].* Would structural holes have a similar behavior pattern in different networks?
3. *Social status [5, 11, 20].* How do different networks satisfy the properties of social status?
4. *"Two-step flow" [18].* How do different networks follow the "two-step flow" of information propagation?

Social Balance Social balance theory suggests that people in a social network tend to form into a balanced network structure. Figure 2 shows such an example to illustrate the structural balance theory over triads, which is the simplest group structure to which balance theory applies. For a triad, the balance theory implies that either all three of these users are friends or only one pair of them are friends. Figure 3 shows the probabilities of balanced triads of the three undirected networks (Epinions, Slashdot, and Mobile). In each network, we compare the probability of balanced triads based on communication links and that based on friendships (or trust relationships). For example, in the Mobile network, the communication links include making a call or sending a message between users. We find it interesting that different networks have very different balance probabilities based on the communication links, e.g., the balance probability in the mobile network is nearly 7 times higher than that of the slashdot network, while based on friendships (or trustful relationships) the three networks have relatively similar balance probabilities (with a maximum of +28% difference).

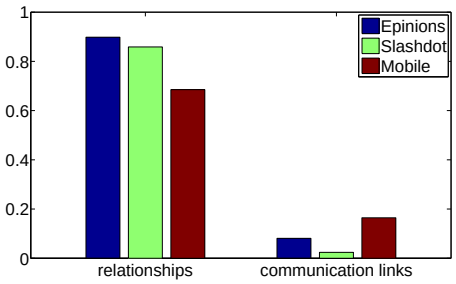


Figure 3: Social balance. Probabilities of balanced triads in different networks based on communication links and friendships (or trustful relationships). Based on communication links, different networks have very different balance probabilities (e.g., the balance probability in the mobile network is nearly 7 times higher than that of the slashdot network). While based on friendships the three networks have a relatively similar probabilities.

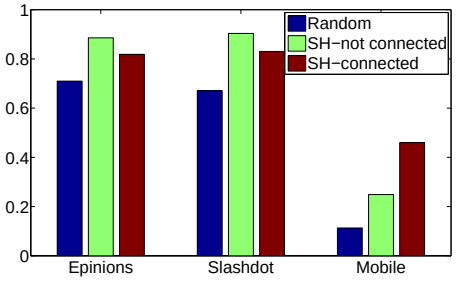


Figure 4: Structural hole. Probabilities that two connected (or disconnected) users (A and B) have the same type of relationship with user C, conditioned on whether user C spans a structural hole or not. It is clear that (1) users are more likely (averagely +70% higher than chance) to have the same type of relationship with C if C spans a structural hole; and (2) disconnected users are more likely than connected users to have the same type of relationship with a user who spans a structural hole (except the mobile network).

Structural Hole Roughly speaking, a person is said to span a *structural hole* in a social network if he or she is linked to people in parts of the network that are otherwise not well connected to one another [3]. Arguments based on structural holes suggest that there is an informational advantage to have friends in a network who do not know each other. A sales manager with a diverse range of connections can be considered as spanning a structural hole, with a number of potentially *weak ties* [9] to individuals in different communities. More generally, we can think about Web sites such as eBay as spanning structural holes, in that they facilitate economic interactions between people who would otherwise not be able to find each other.

Our idea here is to test if a structural hole tends to have the same type of relationship with the other users. We first employ a simple algorithm to identify structural hole users in a network. Following the informal description of structural holes [3], for each node, we count the number of pairs of neighbors who are not directly connected. All users are ranked based on the number of pairs and then top 1% users[4] with the highest numbers are viewed as structural holes in the network. Figure 4 shows the probabilities that two users (A and B) have the same type of relationship with another user (say C), conditioned on whether user C spans a structural hole or not. We have two interesting observations: (1) users are more likely (on average +70% higher than chance) to have the same type of relationship with C if C spans a structural hole; (2) disconnected users are more likely than connected users to have the same type of relationship with a user classified as spanning a structural hole. One exception is the mobile network, where most mobile users in the data set are university students and thus friends frequently communicate with each other.

[4]This is based on the observation that less than 1% of the Twitter users produce 50% of its content [32].

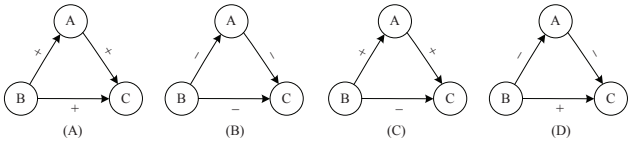


Figure 5: Illustration of status theory. (A) and (B) satisfy the status theory, while (C) and (D) do not satisfy the status theory. Here positive "+" denotes the target node has a higher status than the source node; and negative "-" denotes the target node has a lower status than the source node. In total there are 16 different cases.

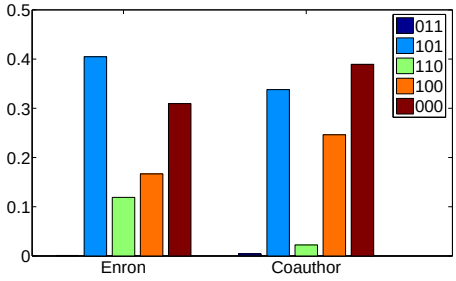


Figure 6: Social status. Distribution of five most frequent formations of triads with social status. Given a triad (A, B, C), let us use 1 to denote the advisor-advisee relationship and 0 colleague relationship. Thus the number 011 to denote A and B are colleagues, B is C's advisor and A is C's advisor.

Social Status Another social psychological theory is the theory of status [5, 11, 20]. This theory is based on the directed relationship network. Suppose each directed relationship labeled by a positive sign "+" or a negative sign "-" (where sign "+"/"-" denotes the target node has a higher/lower status than the source node). Then status theory posits that if, in a triangle on three nodes (also called triad), we take each negative edge, reverse its direction, and flip its sign to positive, then the resulting triangle (with all positive edge signs) should be acyclic. Figure 5 illustrates four examples. The first two triangles satisfy the status ordering and the latter two do not satisfy it. We conducted an analysis on the Coauthor and the Enron networks, where we aim to find directed relationships (advisor-advisee and manager-subordinate). We found nearly 99% of triads in the two networks satisfy the social status theory, which was also validated in [20]. We investigate more by looking at the distribution of different forms of triads in the two networks. Specifically, there are in total 16 different forms of triads [20]. We select five most frequent forms of triads in the two networks. For easy understanding, given a triad (A, B, C), we use 1 to denote the advisor-advisee relationship and 0 colleague relationship, and three consecutive numbers 011 to denote A and B are colleagues, B is C's advisor and A is C's advisor. It is striking that although the two networks (Coauthor and Enron) are totally different, they share a similar distribution on the five frequent forms of triads (as plotted in Figure 6).

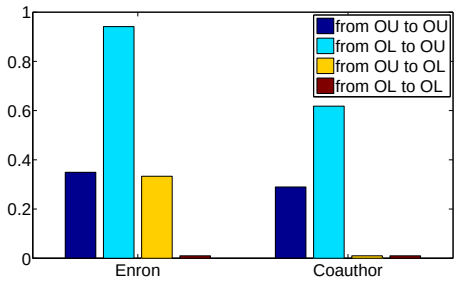


Figure 7: Opinion leader. OL - Opinion leader; OU - Ordinary user. Probability that two types of users have a directed relationship (from higher social status to lower status, i.e., manager-subordinate relationship in Enron and advisor-advisee relationship in Coauthor. It is clear that opinion leaders (detected by PageRank) are more likely to have a higher social-status than ordinary users.

Opinion Leader The two-step flow theory is first introduced in [18] and further elaborated in literature [15, 14]. The theory suggests that ideas (innovations) usually flow first to *opinion leaders*, and then from them to a wider population. In the enterprise email network, for example, managers may act as opinion leaders to help spread information to subordinates.

Our basic idea here is to examine whether "opinion leaders" are more likely to have a higher social status (manager or advisor) than ordinary users. To do this, we first categorize users into two groups (opinion leaders and ordinary users) by PageRank [26][5]. With PageRank, we estimate the importance of each user according to the network structure, and then select as opinion leaders with the top 1% users who have the highest PageRank scores and the rest as ordinary users. Then, we examine the probabilities that two users (A and B) have a directed social relationship (from higher social-status user to lower social-status user) such as advisor-advisee relationship or manager-subordinate relationship. Figure 7 shows some interesting discoveries. First, in both the Enron and Coauthor networks, opinion leaders (detected by PageRank) are more likely (+71%-+84%) to have a higher social status than ordinary users. Second and also more interestingly, in Enron, it is likely that ordinary users have a higher social status than opinion leaders. Its average likelihood is much larger (30 times) than that in the Coauthor network. The reason might be in the enterprise email network (Enron), some managers may be inactive, and most management-related communications were done by their assistants.

Summary According to the statistics above, we have the following intuitions:

1. Probabilities of balanced triads based on communication links are very different in different networks, while the balance probabilities based on friendships (or trustful relationships) are similar with each other.
2. Users are more likely (+25%-+152% higher than chance) to have the same type of relationship with a user who spans a structural hole.
3. Most triads (nearly 99%) satisfy properties of the social status theory. For the five most frequent formations of triads, the Coauthor and the Enron networks have a similar distribution.
4. Opinion leaders are more likely (+71%-+84% higher than chance) to have a higher social status than ordinary users.

[5]PageRank is an algorithm to estimate the importance of each node in a network.

4. MODEL FRAMEWORK

We propose a transfer-based factor graph (TranFG) model for learning and predicting the type of social relationships across network. We first describe the model in the context of a single network, and then explain how to transfer the supervised information provided by one network to another network.

4.1 The Predictive Model

Given a network $G = (V, E^L, E^U, \mathbf{X})$, each relationship (edge) e_i is associated with an attribute vector $\mathbf{x}_i$ and a label y_i indicates the type of the relationship. Let $\mathbf{X} = \{\mathbf{x}_i\}$ and $Y = \{y_i\}$. Then we have the following formulation:

$$P(Y|\mathbf{X}, G) = \frac{P(\mathbf{X}, G|Y)P(Y)}{P(\mathbf{X}, G)} \tag{1}$$

Here, G denotes all forms of network information. This probabilistic formulation indicates that labels of edges depend on not only local attributes associated with each edge, but also the structure of the network. According to Bayes' rule, we have

$$P(Y|\mathbf{X}, G) = \frac{P(\mathbf{X}, G|Y)P(Y)}{P(\mathbf{X}, G)} \propto P(\mathbf{X}|Y) \cdot P(Y|G) \tag{2}$$

where $P(Y|G)$ represents the probability of labels given the structure of the network and $P(\mathbf{X}|Y)$ denotes the probability of generating attributes $\mathbf{X}$ associated to all edges given their labels Y. We assume that the generative probability of attributes given the label of each edge is conditionally independent, thus we have

$$P(Y|\mathbf{X}, G) \propto P(Y|G) \prod_i P(\mathbf{x}_i|y_i) \tag{3}$$

where $P(\mathbf{x}_i|y_i)$ is the probability of generating attributes $\mathbf{x}_i$ given the label y_i. Now, the problem is how to instantiate the probability $P(Y|G)$ and $P(\mathbf{x}_i|y_i)$. In principle, they can be instantiated in different ways, for example according to the Bayesian theory or Markov random fields. In this work, we choose the latter. Thus by the Hammersley-Clifford theorem [12], the two probabilities can be defined as:

$$P(\mathbf{x}_i|y_i) = \frac{1}{Z_1} \exp\{\sum_{j=1}^{d} \alpha_j g_j(x_{ij}, y_i)\} \tag{4}$$

$$P(Y|G) = \frac{1}{Z_2} \exp\{\sum_c \sum_k \mu_k h_k(Y_c)\} \tag{5}$$

where Z_1 and Z_2 are normalization factors. Eq. 4 indicates that we define a feature function $g_j(x_{ij}, y_i)$ for each attribute x_{ij} associated with edge e_i and α_j is the weight of the j^{th} attribute. It can be defined as either a binary function or a real-valued function. For example, for inferring advisor-advisee relationships from the publication network, we can define a real-valued feature function as the difference of years when authors v_i and v_j respectively published his first paper. Such a feature definition is often used in Conditional Random Fields [17] and Maximum Entropy model [23]. Eq. 5 represents that we define a set of correlation feature functions $\{h_k(Y_c)\}_k$ over each clique Y_c in the network. Here μ_k is the weight of the k^{th} correlation feature function. The simplest clique is an edge, thus a feature function $h_k(y_i, y_j)$ can be defined as the correlation between two edges (e_i, e_j), if the two edges share a common end node. We also consider triads as cliques in the TranFG model, in that several social theories we discussed in §3 are based on triads.

If we are given a single network G with labeled information Y, learning the predictive model is to estimate a parameter configuration $\theta = (\{\alpha\}, \{\mu\})$ to maximize the log-likelihood objective function $\mathcal{O}(\theta) = \log P_\theta(Y|\mathbf{X}, G)$, i.e.,

$$\theta^\star = \arg\max \mathcal{O}(\theta) \tag{6}$$

4.2 Learning 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 the two networks. The challenge is then how to bridge the two networks, so that we can transfer the labeled information from the source network to the target network. As the source and target networks may be from arbitrary domains, it is difficult to define correlations between them based on prior knowledge.

To this end, we propose a transfer-based factor graph (TranFG) model. Our idea is based on the fact that the social theories we discussed in §3 are general over all networks. Intuitively, we can leverage the correlation in the extent to which different networks satisfy the different social theories to transfer the knowledge across networks. In particular, for social balance, we define triad based features to denote the proportion of different balanced triangles in a network; for structural hole, we define edge correlation based features, i.e., correlation between two relationships e_i and e_j; for social status, we define features over triads to respectively represent the probabilities of the seven most frequent formations of triads; for opinion leaders, we define features over each edge.

Finally, by incorporating the social theories into our predictive model, we define the following log-likelihood objective function over the source and the target networks:

$$\begin{aligned}\mathcal{O}(\alpha,\beta,\mu) &= \mathcal{O}_S(\alpha,\mu) + \mathcal{O}_T(\beta,\mu)\\ &= \sum_{i=1}^{|V_S|}\sum_{j=1}^{d}\alpha_j g_j(x_{ij}^S, y_i^S) + \sum_{i=1}^{|V_T|}\sum_{j=1}^{d'}\beta_j g'_j(x_{ij}^T, y_i^T)\\ &\quad + \sum_k \mu_k\Big(\sum_{c\in G_S} h_k(Y_c^S) + \sum_{c\in G_T} h_k(Y_c^T)\Big)\\ &\quad - \log Z\end{aligned} \tag{7}$$

where d and d' are numbers of attributes in the source network and the target network respectively. In this objective function, the first term and the second term define the likelihood respectively over the source network and the target network; while the third term defines the likelihood over all common features defined in the two networks. The common feature functions are defined according to the social theories. Such a definition implies that attributes of the two networks can be entirely different as they are optimized with different parameters $\{\alpha\}$ and $\{\beta\}$, while the information transferred from the source network to the target network is the importance of common features that are defined according to the social theories. Finally, we define four (real-valued) balance based features, seven (real-valued) status based features, four (binary) features for opinion leader and six (real-valued) correlation features for structural hole. More details about feature function are given in Appendix.

Model Learning and Inferring The last issue is how to learn the TranFG model and how to infer the type of unknown relationships in the target network. Learning the TranFG model is to estimate a parameter configuration $\theta = (\{\alpha\}, \{\beta\}, \{\mu\})$ to maximize the log-likelihood objective function $\mathcal{O}(\alpha, \beta, \mu)$. We use a gradient decent method (or a Newton-Raphson method) to solve the objective function. We use μ as the example to explain how we learn the parameters. Specifically, we first write the gradient of each μ_k with regard to the objective function:

$$\begin{aligned}\frac{\mathcal{O}(\theta)}{\mu_k} &= \mathbb{E}[h_k(Y_c^S) + h_k(Y_c^T)]\\ &\quad - \mathbb{E}_{P_{\mu_k}(Y_c|\mathbf{X}_S,\mathbf{X}_T,G_S,G_T)}[h_k(Y_c^S) + h_k(Y_c^T)]\end{aligned} \tag{8}$$

where $\mathbb{E}[h_k(Y_c^S) + h_k(Y_c^T)]$ is the expectation of factor function $h_k(Y_c^S) + h_k(Y_c^T)$ given the data distribution (i.e., the average value of the factor function $h_k(Y_c)$ over all triads in the source and the target networks); and the second term $\mathbb{E}_{P_{\mu_k}(Y_c|\mathbf{X}_S,\mathbf{X}_T,G_S,G_T)}[.]$ is the expectation under the distribution $P_{\mu_k}(Y_c|\mathbf{X}_S, \mathbf{X}_T, G_S, G_T)$ given by the estimated model. Similar gradients can be derived for parameter α_j and β_j.

As the graphical structure can be arbitrary and may contain cycles, we use loopy belief propagation (LBP) [24] to approximate the gradients. It is worth noting that to leverage the unlabeled relationships, we need to perform the LBP process twice in each iteration, one time for estimating the marginal distribution of unknown variables $y_i =?$ and the other time for marginal distribution over all cliques. Finally with the gradient, we update each parameter with a learning rate η. The learning algorithm is summarized in Algorithm 1. We see that in the learning process, the algorithm uses an additional loopy belief propagation to infer the label of unknown relationships. After learning, all unknown relationships are assigned with labels that maximize the marginal probabilities.

Input: a source network G_S, a target network G_T, and the learning rate η
Output: estimated parameters $\theta = (\{\alpha\}, \{\beta\}, \{\mu\})$

Initialize $\theta \leftarrow 0$;
Perform statistics according to social theories;
Construct social theories based features $h_k(Y_c)$;
repeat
  Step 1: Perform LBP to calculate marginal distribution of unknown variables in the source network $P(y_i|x_i, G_S)$;
  Step 2: Perform LBP to calculate marginal distribution of unknown variables in the target network $P(y_i|x_i, G_T)$;
  Step 3: Perform LBP to calculate the marginal distribution of clique c, i.e., $P(y_c|\mathbf{X}_c^S, \mathbf{X}_c^T, G_S, G_T)$;
  Step 4: Calculate the gradient of μ_k according to Eq. 8 (for α_j and β_j with a similar formula);
  Step 5: Update parameter θ with the learning rate η:

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta \cdot \frac{\mathcal{O}(\theta)}{\theta}$$

until *Convergence*;

Algorithm 1: Learning algorithm for TranFG.

5. EXPERIMENTAL SETUP

The proposed framework is very general and can be applied to many different networks. For experiments, we consider five different types of networks: Epinions, Slashdot, Mobile, Coauthor, and Enron. On the first three networks (Epinions, Slashdot, and Mobile), our goal is to infer undirected relationships (e.g., friendships), while on the rest two networks (Coauthor and Enron), the goal is to infer directed relationships (e.g., advisor-advisee relationships).

Comparison Methods We compare the following methods for inferring the type of social relationships.

SVM: similar to the logistic regression model [19], SVM uses

Table 2: Performance comparison of different methods for inferring friendships (or trustful relationships). (S) indicates the source network and (T) the target network. For the target network, we use 40% of the labeled data in training and the rest for test.

Data Set	Method	Prec.	Rec.	F1-score
Epinions (S) to Slashdot (T) (40%)	SVM	0.7157	**0.9733**	0.8249
	CRF	0.8919	0.6710	0.7658
	PFG	0.9300	0.6436	0.7607
	TranFG	**0.9414**	0.9446	**0.9430**
Slashdot (S) to Epinions (T) (40%)	SVM	0.9132	**0.9925**	0.9512
	CRF	0.8923	0.9911	0.9393
	PFG	**0.9954**	0.9787	**0.9870**
	TranFG	**0.9954**	0.9787	**0.9870**
Epinions (S) to Mobile (T) (40%)	SVM	0.8983	0.5955	0.7162
	CRF	0.9455	0.5417	0.6887
	PFG	**1.0000**	0.5924	0.7440
	TranFG	0.8239	**0.8344**	**0.8291**
Slashdot (S) to Mobile (T) (40%)	SVM	0.8983	0.5955	0.7162
	CRF	0.9455	0.5417	0.6887
	PFG	**1.0000**	0.5924	0.7440
	TranFG	0.7258	**0.8599**	**0.7872**

attributes associated with each edge as features to train a classification model and then employs the classification model to predict edges' labels in the test data set. For SVM, we employ SVM-light.

CRF: it trains a conditional random field [17] with attributes associated with each edge and correlations between edges.

PFG: the method is also based on CRF, but it employs the unlabeled data to help learn the predictive model. The method is proposed in [29].

TranFG: the proposed approach, which leverages the label information from the source network to help infer the type of relationship in the target network.

We also compare with the method TPFG proposed in [30] for mining advisor-advisee relationships in the publication network. This method is domain-specific and thus we only compare with it on the Coauthor network.

In all experiments, we use the same feature definitions for all methods. On the Coauthor network, we do not consider some domain-specific correlation features[6].

Evaluation Measures To quantitatively evaluate the performance of inferring the type of social relationships, we conducted experiments with different pairs of (source and target) networks, and evaluated the approaches in terms of Precision, Recall and F1-Measure.

All codes were implemented in C++, and all experiments were performed on a PC running Windows 7 with Intel (R) Core (TM) 2 CPU 6600 (2.4GHz and 2.39GHz) and 4GB memory. The efficiency of the proposed TranFG model is acceptable. For example, it took about five minutes to train a TranFG model over the Epinions and the Slashdot networks.

6. RESULTS AND ANALYSIS

In this section, we first evaluate the performance of the proposed approach and the comparison methods. Next, we analyze how social theories can help improve the prediction performance. Finally, we give a qualitative case study to further demonstrate the effectiveness of the proposed approach.

[6]We conducted experiments, but found that those features can lead to overfitting.

Table 3: Performance comparison of different methods for inferring directed relationships (the source end has a higher social status than the target end). (S) indicates the source network and (T) the target network. For the target network, we use 40% of labeled data in training and the rest for test.

Data Set	Method	Prec.	Rec.	F1-score
Coauthor (S) to Enron (T) (40%)	SVM	0.9524	0.5556	0.7018
	CRF	0.9565	0.5366	0.6875
	PFG	**0.9730**	0.6545	0.7826
	TranFG	0.9556	**0.7818**	**0.8600**
Enron (S) to Coauthor (T) (40%)	SVM	0.6910	0.3727	0.4842
	CRF	**1.0000**	0.3043	0.4666
	PFG	0.9916	0.4591	0.6277
	TPFG	0.5936	**0.7611**	0.6669
	TranFG	0.9793	0.5525	**0.7065**

6.1 Performance Analysis

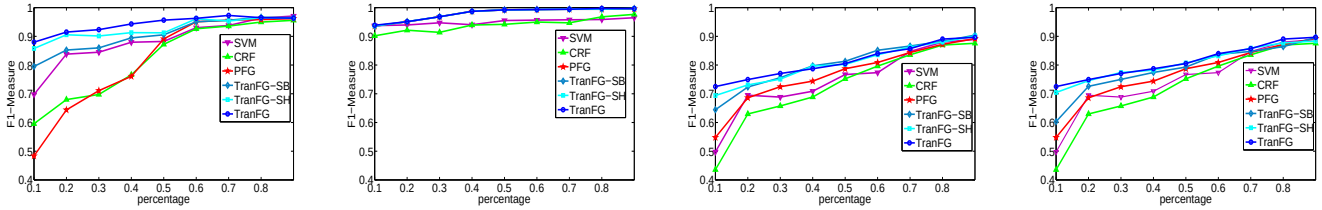
We compare the performance of the four methods for inferring friendships (or trustful relationships) on four pairs of networks: Epinions (S) to Slashdot (T), Slashdot (S) to Epinions (T), Epinions (S) to Mobile (T), and Slashdot (S) to Mobile (T).[7] In all experiments, we use 40% of the labeled data in the target network for training and the rest for test. For transfer, we consider the labeled information in the source network. Table 2 lists the performance of the different methods on the four test cases. Our approach shows better performance than the three alternative methods. We conducted sign tests for each result, which shows that all the improvements of our approach TranFG over the three methods are statistically significant ($p \ll 0.01$).

Table 3 shows the performance of the four methods for inferring directed relationships (the source end has a higher social status than the target end) on two pairs of networks: Coauthor (S) to Enron (T) and Enron (S) to Coauthor (T). We use the same experimental setting as that for inferring friendships on the four pairs of networks, i.e., taking 40% of the labeled data in the target network for training and the rest for test, while for transfer, analogously, we consider the labeled information from the source network. We see that by leveraging the supervised information from the source network, our method clearly improves the performance (about 15% by F1-score on Enron and 10% on Coauthor).

The method PFG can be viewed as a non-transferable counterpart of our method, which does not consider the labeled information from the source network. From both Table 2 and Table 3, we can see that with the transferred information, our method can clearly improve the relationship categorization performance. Another phenomenon is that PFG has a better performance than the other two methods (SVM and CRF) in most cases. PFG could leverage the unlabeled information in the target network, thus improves the performance. The only exception is the case of Epinions (S) to Slashdot (T), where it seems that users in Slashdot have a relatively consistent pattern and merely with some general features such as in-degree, out-degree, and number of common neighbors, a classification based method (SVM) can achieve very high performance.

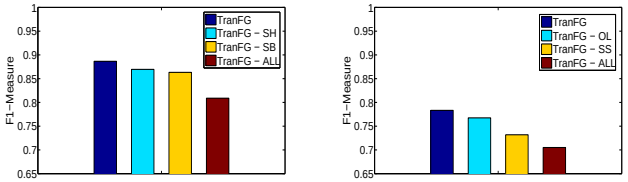
Factor contribution analysis We now analyze how different social theories (social balance, social status, structural hole, and two-

[7]We did try to use Mobile as the source network and Slashdot/Epinions as the target network. However as the size of Mobile is much smaller than the other two networks, the performance was considerably worse.



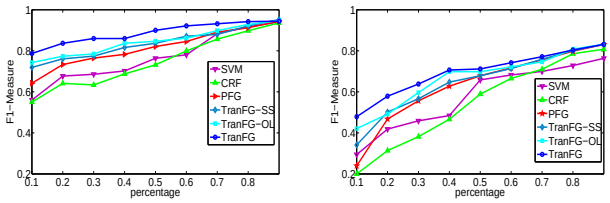
(a) Epinions-to-Slashdot (b) Slashdot-to-Epinions (c) Epinions-to-Mobile (d) Slashdot-to-Mobile

Figure 9: Performance of inferring friendships with and w/o the balance based transfer by varying the percent of labeled data in the target network.



(a) Friendship (b) Directed

Figure 8: Factor contribution analysis. TranFG-SH denotes our TranFG model by ignoring the structural hole based transfer. TranFG-SB stands for ignoring the structural balance based transfer. TranFG-OL stands for ignoring the opinion leader based transfer and TranFG-SS stands for ignoring social status based transfer.



(a) Coauthor-to-Enron (b) Enron-to-Coauthor

Figure 10: Performance of inferring directed relationship with and w/o the status based transfer by varying the percent of labeled data in the target network.

step flow (opinion leader)) can help infer social ties. For inferring friendships, we consider social balance (SB) and structural hole (SH) based transfer and for inferring directed friendships, we consider social status (SS) and opinion leader (OL) based transfer. Here we examine the contribution of the different factors defined in our TranFG model. Figure 8 shows the average F1-Measure score over the different networks, obtained by the TranFG model for inferring friendships and directed relationships. In particular, TranFG-SB represents that we remove social balance based transfer features from our model and TranFG-All denotes that we remove all the transfer features. It can be clearly observed that the performance drops when ignoring each of the factors. We can also see that for inferring friendships the social balance is a bit more useful than structural hole, and for inferring directed relationships the social status factor is more important than the factor of opinion leader. The analysis also confirms that our method works well (further improvement is obtained) when combining different social theories.

Social balance and structural hole based transfer. We present an in-depth analysis on how the social balance and structural hole based transfer can help by varying the percent of labeled training data in the target network. We see that in all cases except Slashdot-to-Epinions, clear improvements can be obtained by using the social balance and structural hole based transfer, when the labeled data in the target network is limited ($\leq 50\%$). Indeed, in some case such as Epinions-to-Slashdot, merely with 10% of the labeled relationships in Slashdot, our method can obtain a good performance (88% by F1-score). Without transfer, the best performance is only 70% (obtained by SVM). We also find that structural balance based transfer is more helpful than structural hold based transfer for inferring friendships in most cases with various percents of labeled relationships. This result is consistent with that obtained in the factor contribution analysis.

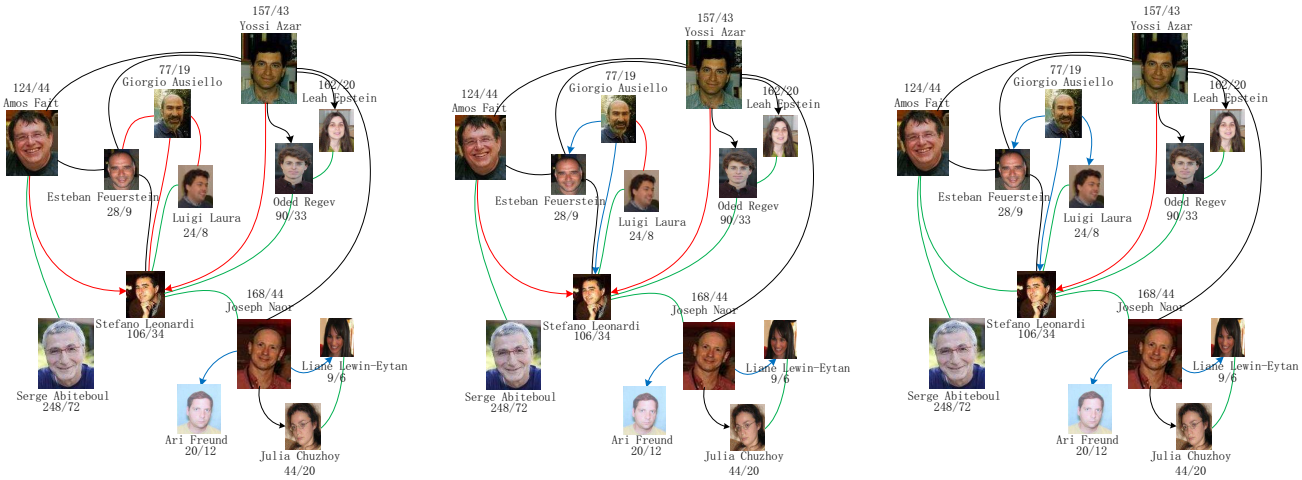
A different phenomenon is found in the case of Slashdot-to-Epinions, where all methods can obtain a F1-score of 94% with only 10% of the labeled data. The knowledge transfer seems not helpful. By a careful investigation, we found simply with those features (Cf. Appendix for details) defined on the edges, we could achieve a high performance (about 90%). The structure information indeed helps, but the gained improvement is limited.

Social status and opinion leader based transfer. Figure 10 shows an analysis for inferring directed relationships on the two cases (Enron-to-Coauthor and Coauthor-to-Enron). Here, we focus on testing how social status and opinion leader based transfer can help infer the type of relationships by varying the percent of labeled relationships in the target network. In both cases (Coauthor-to-Enron and Enron-to-Coauthor), the TranFG model achieves consistent improvements. For example, when there is only 10% of labeled advisor-advisee relationships in the Coauthor network, without considering the status and opinion leader based transfer, the F1-score is only 24%. By leveraging the status and opinion leader based transfer from the email network (Enron), the score is doubled (47%). Moreover, we find that the social status based transfer is more helpful than the opinion leader based transfer with various percents of the labeled data.

6.2 Case Study

Now we present a case study to demonstrate the effectiveness of the proposed model. Figure 11 shows an example generated from our experiments. It represents a portion of the Coauthor network. Black edges and arrows respectively denote labeled colleague relationships and advisor-advisee relationships in the training data. Colored arrows and edges indicate advisor-advisee and colleagues relationships detected by three methods: SVM, PFG and TranFG, with red color indicating mistake ones. The numbers associated with each author respectively denote the number of papers and the score of h-index.

We investigate more by looking at a specific example. SVM mistakenly classifies three advisor-advisee relationship and two col-



(a) SVM (b) PFG (c) Our approach (TranFG)

Figure 11: Case study. Illustration of inferring advisor-advisee relationships on the Coauthor network. Directed edges indicate advisor relationships, and undirected ones indicate coauthor relationships. Black edges indicate labeled data. Red colored edges indicates wrong predictions.

league relationships. SVM trains a local classification model without considering the network information. PFG considers the network information as well as the unlabeled data, thus obtains a better result. Our proposed TranFG model further corrects two mistakes ("Fait-Leonardi" and "Ausiello-Laura") by leveraging properties of social status and opinion leader. For example, the results obtained by PFG among "Azar", "Amos" and "Leonardi" form a triad of ("011"). Although it satisfies the property of social status, the probability of such triad is much lower $(0.4\%$ vs. $24.6\%)$ than the form ("100"). However, the limitation of the training data leads PFG to result in a bias mistake $(5.8\%$ vs. $12.6\%)$. TranFG smoothes the inferring results by transferring knowledge from the source (Enron) network.

7. RELATED WORK

Inferring social ties is an important problem in social network analysis. Liben-Nowell et al. [21] present a unsupervised method for link prediction. Xiang et al. [33] develop a latent variable model to estimate relationship strength from interaction activity and user similarity. Backstrom et al. [2] propose a supervised random walk algorithm to estimate the strength of social links. Leskovec et al. [19] employ a logistic regression model to predict positive and negative links in online social networks. Hopcroft et al. [13] study the extent to which the formation of a reciprocal relationship can be predicted in a dynamic network. However, most existing works focus on predicting and recommending unknown links in social networks, but ignore the type of relationships.

Recently, there are several works on inferring the meanings of social relationships. Diehl et al. [6] try to identify the manager-subordinate relationships by learning a ranking function. Wang et al. [30] propose an unsupervised probabilistic model for mining the advisor-advisee relationships from the publication network. Crandall et al. [4] investigate the problem of inferring friendship between people from co-occurrence in time and space. Eagle et al. [7] present several patterns discovered in mobile phone data, and try to use these patterns to infer the friendship network. However, these algorithms mainly focus on a specific domain, while our model is general and can be applied to different domains. More importantly, our work takes the first step to incorporate social theories for inferring social ties across heterogeneous networks.

Our work is related with link prediction, which is one of the core tasks in social networks. Existing work on link prediction can be broadly grouped into two categories based on the learning methods employed: unsupervised link prediction and supervised link prediction. Unsupervised link predictions usually assign scores to potential links based on the intuition - the more similar the pair of users are, the more likely they are linked. Various similarity measures of users are considered, such as the Adamic and Adar measure [1], the preferential attachment [25], and the Katz measure [16]. A survey of unsupervised link prediction can be found in [21]. Recently, [22] designs a flow based method for link prediction. There are also a few works which employ supervised approaches to predict links in social networks, such as [31, 2, 19]. The main difference between existing work on link prediction and our effort lies in that existing work mainly focuses on specific domains, while our proposed model combines social theories (such as structural balance, structural hole, and social status) into a transfer learning framework and can be easily applied to different domains.

8. CONCLUSION

In this paper, we study the novel problem of inferring social ties across heterogeneous networks. We precisely define the problem and propose a transfer-based factor graph (TranFG) model. The model incorporates social theories into a semi-supervised learning framework, which is used to transfer supervised information from the source network to help infer social ties in the target network. We evaluate the proposed model on five different genres of networks. We show that the proposed model can significantly improve the performance for inferring social ties across different networks comparing with several alternative methods. Our study also reveals several interesting phenomena.

The general problem of inferring social ties represents a new and interesting research direction in social network analysis. There are many potential future directions of this work. First, some other social theories can be further explored and validated for analyzing the formation of different types of social relationships. Next, it is also interesting to study how to further correct the inferring mistakes by involving users into the learning process (e.g., via active learning). Another potential issue is to validate the proposed model on some other social networks.

9. REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *SOCIAL NETWORKS*, 25:211–230, 2001.

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

[3] R. S. Burt. *Structural Holes : The Social Structure of Competition*. Cambridge, Mass.: Harvard University Press, 1995.

[4] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 107:22436–22441, Dec. 2010.

[5] J. A. Davis and S. Leinhardt. The structure of positive interpersonal relations in small groups. In J. Berger, editor, *Sociological Theories in Progress*, volume 2, pages 218–251. Houghton Mifflin, 1972.

[6] C. P. Diehl, G. Namata, and L. Getoor. Relationship identification for social network discovery. In *AAAI*, pages 546–552, 2007.

[7] N. Eagle, A. S. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. *PNAS*, 106(36), 2009.

[8] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.

[9] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[10] R. Grob, M. Kuhn, R. Wattenhofer, and M. Wirz. Cluestr: Mobile social networking for enhanced group communication. In *GROUP'09*, 2009.

[11] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW'04*, pages 403–412, 2004.

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

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

[14] E. Katz. The two-step flow of communication: an up-to-date report of an hypothesis. *In Enis and Cox(eds.), Marketing Classics*, pages 175–193, 1973.

[15] E. Katz and P. F. Lazarsfeld. *Personal Influence*. The Free Press, New York, USA, 1955.

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

[17] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'01*, pages 282–289, 2001.

[18] P. F. Lazarsfeld, B. Berelson, and H. Gaudet. *The people's choice: How the voter makes up his mind in a presidential campaign*. Columbia University Press, New York, USA, 1944.

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

[20] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *CHI'10*, pages 1361–1370, 2010.

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

[22] R. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *KDD'10*, pages 243–252, 2010.

[23] A. McCallum, D. Freitag, and F. C. N. Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML'00*, pages 591–598, 2000.

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

[25] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64(2):025102, 2001.

[26] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University, 1999.

[27] M. Roth, A. Ben-David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, and R. Merom. Suggesting friends using the implicit social graph. In *KDD'10*, 2010.

[28] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998, 2008.

[29] W. Tang, H. Zhuang, and J. Tang. Learning to infer social relationships in large networks. In *ECML/PKDD'11*, 2011.

[30] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *KDD'10*, pages 203–212, 2010.

[31] C. Wang, V. Satuluri, and S. Parthasarathy. Local probabilistic models for link prediction. In *ICDM'07*, pages 322–331, 2007.

[32] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *WWW'11*, 2011.

[33] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *WWW'10*, pages 981–990, 2010.

Appendix

There are two categories of features. The first category includes local features defined for each specific network, and the second includes transfer features defined based on the social theories. Tables 4-7 give a summary of local feature definitions for the five networks. For a more detailed description of the feature definitions, please refer to literature [6, 7, 19, 29].

For the transfer features, in Epinions, Slashdot and Mobile, we define four (real-valued) balance triad based features and six (real-valued) structural hole based features. In the Coauthor and Enron, we define seven (real-valued) social status based features and four (binary) opinion leader based features.

Table 4: Features defined for edge (v_i, v_j) in Epinions/Slashdot.

Feature	Description
in-degree	$d_{in}(v_i)$, $d_{in}(v_j)$
out-degree	$d_{out}(v_i)$, $d_{out}(v_j)$
total-degree	$d_{in}(v_i) + d_{out}(v_i)$, $d_{in}(v_j) + d_{out}(v_j)$
common neighbors	the total number of common neighbors of v_i and v_j in an undirected sense.

Table 5: Features defined for edge (v_i, v_j) in Mobile.

Feature	Description
Total Proximity	Number of proximity events between a and b
In-Role	Number of proximity events at working place in day time from Monday to Friday
Extra-Role	Number of proximity events at home or elsewhere at night of weekends
Total Communication	Number of communication logs between a and b
Night Call Ratio	The ratio of communication logs at night

Table 6: Features defined for edge (v_i, v_j) in Coauthor. P_i denotes a set of papers published by author v_i.

Feature	Description
paper count	$\|P_i\|$, $\|P_j\|$
paper ratio	$\|P_i\|/\|P_j\|$
coauthor ratio	$\|P_i \cap P_j\|/\|P_i\|$, $\|P_i \cap P_j\|/\|P_j\|$
conference coverage	The proportion of the conferences which both v_i and v_j attended among conferences v_j attended.
first-pub-year diff	The difference in year of the first earliest publication of v_i and v_j.

Table 7: Features defined for edge (v_i, v_j) in Enron.

From	Sent-To + CC	From	Sent-To + CC
v_i	v_j	v_j	v_i
v_i	v_k and not v_j	v_j	v_k and not v_i
v_k	v_i and not v_j	v_k	v_j and not v_i
v_k	v_i and v_j		