Link Prediction with Signed Latent Factors in Signed Social Networks

Pinghua Xu*
School of Computer Science
Wuhan University, China
xupinghua@whu.edu.cn

Wenbin Hu†
School of Computer Science
Wuhan University, China
hwb@whu.edu.cn

Jia Wu
Department of Computing
Macquarie University, Australia
jia.wu@mq.edu.au

Bo Du†
School of Computer Science
Wuhan University, China
remoteking@whu.edu.cn

ABSTRACT

Link prediction in signed social networks is an important and challenging problem in social network analysis. To produce the most accurate prediction results, two questions must be answered: (1) Which unconnected node pairs are likely to be connected by a link in future? (2) What will the signs of the new links be? These questions are challenging, and current research seldom well solves both issues simultaneously. Additionally, neutral social relationships, which are common in many social networks can affect the accuracy of link prediction. Yet neutral links are not considered in most existing methods. Hence, in this paper, we propose a **s**igned **l**atent **f**actor (SLF) model that answers both these questions and, additionally, considers four types of relationships: positive, negative, neutral and no relationship at all. The model links social relationships of different types to the comprehensive, but opposite, effects of positive and negative SLFs. The SLF vectors for each node are learned by minimizing a negative log-likelihood objective function. Experiments on four real-world signed social networks support the efficacy of the proposed model.

CCS CONCEPTS

• **Information systems → Social networks**; *Data mining*.

KEYWORDS

link prediction; signed latent factor; signed social network

ACM Reference Format:
Pinghua Xu, Wenbin Hu, Jia Wu, and Bo Du. 2019. Link Prediction with Signed Latent Factors in Signed Social Networks. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330850

1 INTRODUCTION

Signed social networks reflect complex relationships in real-world social sites better than unsigned social networks because connected node pairs (i.e., social relatonships or links) can be labeled as positive, negative, or neutral [6, 21, 27, 29]. Further, any two nodes that are not connected indicate the absence of a social relationship. For example, in an election network, a voter may cast a positive, negative, or neutral vote, or could choose not to vote at all [5]. Given the many pratical uses for this kind of network structure, this paper focuses on link prediction in signed social networks.

The most important characteristic of a signed social network is the variety of the social relationships it contains. Hence, the two corresponding challenges that need to be emphasized when solving the link prediction problems are:

- **Q1**: Which unconnected node pairs are likely to be connected by a link in future?
- **Q2**: What will the signs of those new links be?

Many existing studies on signed social networks are based on the social theories (e.g., structural balance theory [16]), which transform a link prediction problem into a adjacency matrix completion problem [16, 30]. Since many signed social networks are consistent with the social theories, these approaches are good at accurately predicting the signs of the new links (Q2). However, the resulting predictions are based on an unrealistic assumption - that we already know which nodes will be connected. Hence, approaches based on social theories typically cannot predict which two nodes will form a new connection (Q1). Unlike signed social networks, predicting new links in an unsigned network is an inherent task [1, 2, 4, 8, 31]. The methods designed for unsigned social networks can also predict the signs by converting the original network to subnetworks with only one type of link. However, incomplete networks with only one type of link do not contain the important topology information that would have been available in the original networks. Therefore, those approaches have the opposite problem - they are good at predicting new links (Q1) but not good at predicting their signs (Q2). Recently, some scholars have begun to more fully consider

[*] Also with Department of Computing, Macquarie University, Australia.
[†] Corresponding authors, and Wenbin Hu is also with Shenzhen Research Institute, Wuhan University, China.

KDD ’19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330850

the emergence of new links in signed social networks [18, 19, 26]. The feature-based models mainly rely on the cycle structures [6], especially the triad structures. In a dense network, cycles are abundant, which makes training an effective prediction model relatively easy. However, when there are not enough cycles, as with a sparse network, performance suffers. Unfortunately, most signed social networks are very sparse. So, although these models do address both Q1 and Q2 simultaneously, the predictions are not very accurate.

The **s**igned **l**atent **f**actor (SLF) model presented in this paper has been designed to overcome all these challenges - predicting new links, predicting the signs of links, and making accurate predictions in networks of any sparsity. Consider, for example, a simple social site where users tend to like other users that have a sense of humor and dislike the users that are boring. In this example, "humorous" is a positive characteristic, i.e., positive SLF. Therefore, if user v has a sense of humor, user u is likely to give positive feedback to user v, prompting a positive social relationship between the two. The level of positivity is related to user u's level of interest and v's composition of humour. Moreover, the effect and the amount of the positive feedback is positively related. Similarly, the "boring" characteristic is a negative SLF, which will have the opposite effect on the relationships. Note, however, that although the positive and negative SLFs in this example represent personal characteristics, SLFs do not need to have an explicit meaning like the latent factors in traditional models [15].

Thus, we have designed two types of SLFs to work with our model - one positive and one negative. Each yields the opposite effect on formation of social relationships. The two types of SLFs are mapped onto two independent SLF spaces, where each dimension represents a SLF. Further, within each space, each node is characterized by two SLF vectors, which represent the interests and compositions, respectively. These vectors are used to capture the directionality of the relationship. The overall approach is encapsulated in a 3-step procedure specifically-designed for the SLF model. Nonlinear factors and sociological meanings are used to produce four different types SLF scores for a node pair, each corresponding to the four types of social relationships considered in the model.

These four relationships are positive, negative, neutral, and no relationship at all. Positive and negative relationships are the most commonly considered. However, neutral social relationships can also reflect the propensity of one node to connect with another [3, 24, 25]. For instance, in the election network [5], we generally believe that an active voter who has cast many neutral votes will continue to vote that way in future rather than choose not to vote. If considered, this information would affect the prediction results. However, most existing approaches either ignore the neutral social relationships or consolidate them into the "no relationship" basket. The usual consequence is that a model will underestimate on the propensity of a node to connect to another. In contrast, the SLF model presented in this paper makes good use of the neutral relationships. The SLF vectors are learned by minimizing a negative log-likelihood objective function, which not only considers the positive and negative relationships but also neutral and absent relationships. Further, the scores assigned to neutral links have a reasonable and sociological meaning. Null relationships, where no relationship exists, are another important consideration - especially since most social networks, even dense ones, contain a large number of null relationships. Optimizing null relationships are time consuming but many of these relationships are uninformative at the process of learning the SLF vectors. Therefore, we have incorporated a node pair sampling process into the optimization procedure.

Table 1: Common notations.

Notation	Description
G	A signed social network
V	The set of nodes
E^p	The set of positive links
E^n	The set of negative links
E^{ne}	The set of neutral links
E^{non}	The set of unconnected node pairs
U^{out}	Positive outward SLF vectors
U^{in}	Positive inward SLF vectors
W^{out}	Negative outward SLF vectors
W^{in}	Negative inward SLF vectors
k_1	Dimensions of the positive SLF space
k_2	Dimensions of the negative SLF space
n	Sample size of null social relationships

Our experiments involve link prediction tasks on four real-world signed social networks. The results demonstrate the superiority of our proposed SLF model over the state-of-the-art methods.

The rest of this paper is organized as follows. The preliminaries are explained in Section 2. The proposed SLF model [1] is presented in Section 3. SLF's link prediction precedure is outlined in Section 4. Section 5 details the experiments, followed by the conclusion in Section 6.

2 PRELIMINARIES

This section begins with the problem formulation and a brief explaination of signed latent factors. Table 1 summarizes the common notations used in the following sections.

2.1 Problem Formulation

Consider a signed social network $G = (V, E^p, E^n, E^{ne})$, where V is the set of users and E^p, E^n, and E^{ne} are the sets of positive, negative, and neutral links. Unconnected node pairs are defined as null relationships, denoted as E^{non}.

The link prediction problem to be addressed is to predict if the node pairs in E^{non} will be connected by a positive, negative, or neutral link, or remain unconnected in future.

2.2 Signed Latent Factor

Formally, the latent factors in a social network are the independent unobserved variables which can be used to describe the observed social relationships [13]. Since a signed social network inherently contains social relationships that are more complex, which cannot be well captured by traditional latent factor models, we have designed two types of SLFs - one positive and one negative - each with the corresponding, but opposite, effect on a social relationship. Each node

[1] The implementation is available at: https://github.com/WHU-SNA/SLF

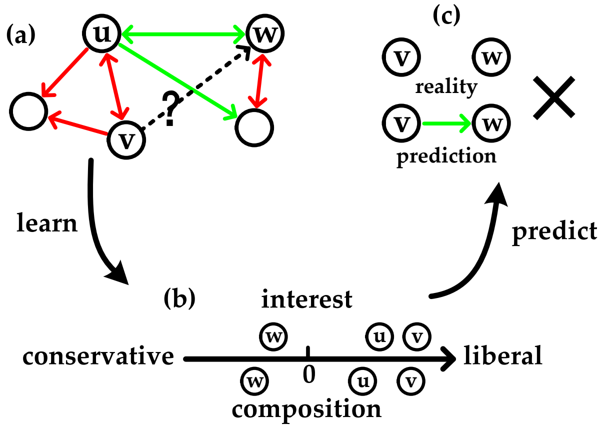

Figure 1: (a) is a political network where the red and green solid lines represent positive and negative social relationships, respectively. (b) is a simplified one-dimensional latent factor space. Each node corresponds to two positions in the latent factor space - interest and composition. Determining a node pair's relationship using a traditional model - i.e., by the positive and negative values of the latent factor vectors - will result in a negative prediction for node pair (v, w), which is wrong!

is mapped onto two independent SLF spaces. Within each space, each node is characterized by two SLF vectors. Hence, each node is characterized by a total of four SLF vectors, two in the positive space and two in the negative. A node's positive outward/inward SLF vector represents its personal interests/compositions toward positive SLFs. Similarly, a node's negative outward/inward SLF vector represents its personal interests/compositions toward negative SLFs.

In traditional latent factor models [10, 15, 17, 22, 23], using the positive and negative values of latent factor vectors is a natural way to represent a node's relationships toward positive and negative SLFs - for example, a bias toward either liberal or conservative politics. As such, if a node has an interest in the positive parts of a latent factor, it must have a negative social relationships to the nodes with a composition to the negative parts of that latent factor. However, this promise is often unreasonable in the real world. As shown in Figure 1, users u and v are liberals who trust other liberals, but only user u distrusts the conservatives; user v is ambivalent. A traditional latent factor model would predict user v to be distrustful of the conservative user w. Whereas, our SLF model is able to distinguish between the different behavior patterns of u and v using the two different SLFs. Hence, the resulting predictions would be more accurate. This conjecture is indeed supported by the results of our experiments (Section 5.2).

However, as our model has two independent SLF spaces, a new approach is needed to score each node pair against the relevant factors. This approach is presented in the next section.

3 THE SLF MODEL

The most important procedure in our SLF model is the method of calculating the node pair scores.

Here, a comparison between a trditional latent factor model and the proposed SLF model is useful. In this rudimentary example, assume that there are two types of social relationships - i.e., in some way related or not related. Given the node pair (u, v), a traditional model simply uses the inner product of the outward latent factor vector of u and the inward latent factor vector of v as the score for whether a social relationship will form between the nodes. However, traditional models do not contain a nonlinear factor, so their fitting performance is limited.

Obviously, the same approach would not work with our SLF model as we have two independent SLF spaces and must consider four types of social relationships. Hence, we devised a 3-step procedure (see Figure 2) to calculate the four types of scores:

- **Step 1.** Given the node pair (u, v), we first calculate the inner product of U_u^{out} and U_v^{in} and call the result positive feedback F_{uv}^+. Then we calculate the inner product of W_u^{out} and W_v^{in} and call the result negative feedback F_{uv}^-.
- **Step 2.** Next, we quantify the effects of the positive and negative feedbacks on the social relationship, formulated as $f_a(U_u^{out}U_v^{in})$ and $f_a(W_u^{out}W_v^{in})$ respectively, where $f_a(x) = \frac{p_0 exp(x)}{1+p_0(exp(x)-1)}$ is a logistic activation function, and p_0 represents the effect of no feedback.
- **Step 3.** The scores for forming a positive, negative, neutral or null relationship are defined as $F_{uv}^+(1-F_{uv}^-)$, $(1-F_{uv}^+)F_{uv}^-$, $F_{uv}^+F_{uv}^-$ and $(1-F_{uv}^+)(1-F_{uv}^-)$ respectively.

The logistic activation function in Step 2 and the meanings of the scores in Step 3 are key.

- The activation function $f_a(x)$ constructs a nonlinear mapping of the positive and negative feedback and its effects on the relationships. It also normalizes the effects to the interval $[p_0, 1)$. As the feedback associated with a sign incrementally increases, the value of its effect initially grows slowly, then quickly, and eventually forms an S-curve, which is reflective of some real-world scenarios. Hence, $f_a(x)$ offers a common criterion for measuring the effect of different types of feedback, and the inherent nonlinear factor has the potential to improve the fitting capacity of the SLF model.
- The formation of a social relationship can be seen as the result of comprehensive interplays between positive and negative feedback. Hence, the score for forming a positive social relationship reflects a sociological perspective. Specifically, if a person u very much appreciates v, and only sees a small amount of bad in v (i.e., the likelihood $F_{uv}^+(1-F_{uv}^-)$ is large), we tend to believe that the social relationship between u and v is positive. The score for forming a negative social relationship has a similar meaning. In addition, small values of F_{uv}^+ and F_{uv}^- indicate that u is not interested in v either way. If the value of $(1-F_{uv}^+)(1-F_{uv}^-)$ is large, we tend to believe that u and v do not have a social relationship, i.e., they have a null relationship. Conversely, a large value of $F_{uv}^+F_{uv}^-$ indicates that u likes v in many aspects, but also dislikes v a great deal in other respects. Notwithstanding the

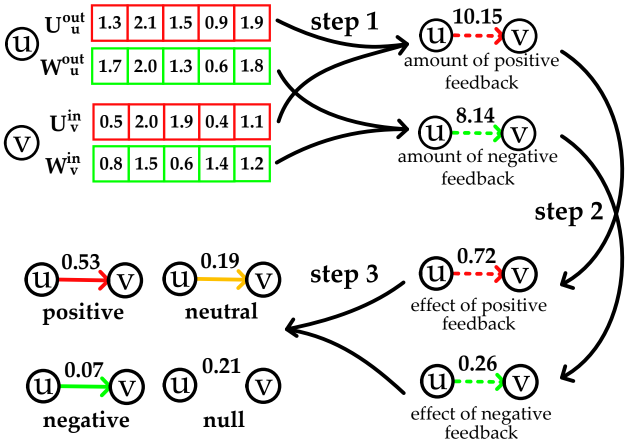
u
U_u^{out} 1.3 2.1 1.5 0.9 1.9
W_u^{out} 1.7 2.0 1.3 0.6 1.8
v
U_v^{in} 0.5 2.0 1.9 0.4 1.1
W_v^{in} 0.8 1.5 0.6 1.4 1.2
step 1
10.15
amount of positive feedback
8.14
amount of negative feedback
step 2
0.72
effect of positive feedback
0.26
effect of negative feedback
step 3
0.53
positive
0.19
neutral
0.07
negative
0.21
null

Figure 2: The 3-step procedure for calculating the four types of node pair scores.

love-hate connotation, this type of relationship tends to be neutral.

The formal definition of the SLF model is as follows:

Definition 1 (Signed latent factor model). Let V be the set of nodes in a signed social network. Given the SLF vectors U^{out}, U^{in}, W^{out} and W^{in}, the scores for node u to generate a positive, negative, neutral or null relationship to node v are $f_a(U_u^{out}U_v^{in})(1 - f_a(W_u^{out}W_v^{in}))$, $(1 - f_a(U_u^{out}U_v^{in}))f_a(W_u^{out}W_v^{in})$, $f_a(U_u^{out}U_v^{in})$ $f_a(W_u^{out}W_v^{in})$ and $(1 - f_a(U_u^{out}U_v^{in}))(1 - f_a(W_u^{out}W_v^{in}))$ respectively.

In summary, the SLF model contains two types of SLFs, and each node is characterized by four SLF vectors. The scores that determine whether and which type of relationship will be formed between two nodes are calculated through a 3-step procedure. Each type of score has a corresponding sociological meaning.

4 LINK PREDICTION WITH THE SLF MODEL

This section explains the procedure for solving a link prediction problem in a signed social network using the SLF model. The most important step is learning SLF vectors from the given signed social network.

4.1 The Link Prediction Procedure

The link prediction procedure (see Figure 3) has five steps as follows:

- **Step 1.** Learn the SLF vectors for each node from the given signed social network G.
- **Step 2.** Concatenate the SLF vectors for u, U_u^{out}, U_u^{in}, W_u^{out} and W_u^{in}, to compose a node feature f_u for node u. That is to say, $f_u = (U_u^{out}, U_u^{in}, W_u^{out}, W_u^{in})$
- **Step 3.** Combine the node features for u and v to compose a node pair feature $f_{(u,v)}$ for the node pair (u, v).
- **Step 4.** Use the labels and features of the node pairs to train a classification model.
- **Step 5.** Use the classification model and the node pair features to make predictions.

The node pair feature in Step 3 could be constructed in many different ways. Here we consider the most common four operators (see Table 2). The impact of each operator on prediction performance is discussed in Section 5. The classification model we chose for Step 5 is the simple and widely used logistic regression model.

Table 2: Operators for constructing node pair features.

Operator	Definition
Average (Avg)	$f_{(u,v)} = \frac{1}{2}(f_u + f_v)$
Concatenate (Con)	$f_{(u,v)} = (f_u, f_v)$
L1_weight (L1)	$f_{(u,v)} = \lvert f_u - f_v \rvert$
L2_weight (L2)	$f_{(u,v)} = \lvert f_u - f_v \rvert^2$

4.2 Learning SLF Vectors

The most important step of the procedure is Step 1: learning the SLF vectors for each node from the signed social network. The maximum likelihood formulation that links the SLF vectors to the social relationships takes the form of a negative log-likelihood objective function, defined as follows:

$$\begin{aligned} L = &- \textstyle\sum_{(u,v)\in E^p} log(F_{uv}^+(1 - F_{uv}^-)) \\ &- \textstyle\sum_{(u,v)\in E^n} log((1 - F_{uv}^+)F_{uv}^-) \\ &- \textstyle\sum_{(u,v)\in E^{ne}} log(F_{uv}^+ F_{uv}^-) \\ &- \textstyle\sum_{(u,v)\in E^{non}} log((1 - F_{uv}^+)(1 - F_{uv}^-)) \end{aligned} \tag{1}$$

where $F_{uv}^+ = f_a(U_u^{out}U_v^{in})$, $F_{uv}^- = f_a(W_u^{out}W_v^{in})$. And the logistic activation function f_a is formulated as $f_a(x) = \frac{p_0 exp(x)}{1+p_0(exp(x)-1)}$.

Equation 1 consists of four components, which correspond to the positive, negative, neutral and null relationships. To minimize the objective function, we use the coordinate descent method [11, 20], where the SLF vectors of a node are updated while fixing all other SLF vectors in each iteration. Hence, the problem becomes a convex optimization problem, and the objective function becomes:

$$\begin{aligned} L_{(u)} = &- \textstyle\sum_{v\in N_{out}^p(u)} log(F_{uv}^+(1 - F_{uv}^-)) \\ &- \textstyle\sum_{v\in N_{in}^p(u)} log(F_{vu}^+(1 - F_{vu}^-)) - \textstyle\sum_{v\in N_{out}^n(u)} log((1 - F_{uv}^+)F_{uv}^-) \\ &- \textstyle\sum_{v\in N_{in}^n(u)} log((1 - F_{vu}^+)F_{vu}^-) - \textstyle\sum_{v\in N_{out}^{ne}(u)} log(F_{uv}^+ F_{uv}^-) \\ &- \textstyle\sum_{v\in N_{in}^{ne}(u)} log(F_{vu}^+ F_{vu}^-) - \textstyle\sum_{v\in N_{out}^{non}(u)} log((1 - F_{uv}^+)(1 - F_{uv}^-)) \\ &- \textstyle\sum_{v\in N_{in}^{non}(u)} log((1 - F_{vu}^+)(1 - F_{vu}^-)) \end{aligned} \tag{2}$$

where $N_{out}^p(u)$, $N_{out}^n(u)$, $N_{out}^{ne}(u)$ and $N_{out}^{non}(u)$ denote the successors of u with positive, negative, neutral and null relationships. Similarly, $N_{in}^p(u)$, $N_{in}^n(u)$, $N_{in}^{ne}(u)$ and $N_{in}^{non}(u)$ denote the predecessors of u with different social relationships.

The derivate of Equation 2 is as follows:

$$\begin{aligned} \frac{\partial L_{(u)}}{\partial U_u^{out}} = &- \textstyle\sum_{v\in N_{out}^p(u)\cup N_{out}^{ne}(u)}((1 - F_{uv}^+)U_v^{in}) \\ &+ \textstyle\sum_{v\in N_{out}^n(u)\cup N_{out}^{non}(u)} F_{uv}^+ U_v^{in} \end{aligned} \tag{3}$$

$$\begin{aligned} \frac{\partial L_{(u)}}{\partial U_u^{in}} = &- \textstyle\sum_{v\in N_{in}^p(u)\cup N_{in}^{ne}(u)}((1 - F_{vu}^+)U_v^{out}) \\ &+ \textstyle\sum_{v\in N_{in}^n(u)\cup N_{in}^{non}(u)} F_{vu}^+ U_v^{out} \end{aligned} \tag{4}$$

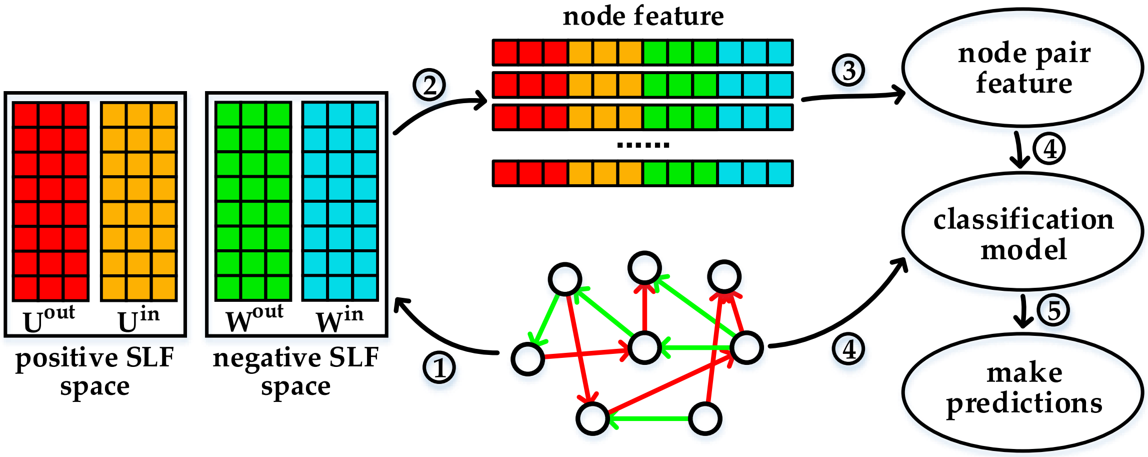

Figure 3: The SLF model based procedure for solving the link prediction problem in a signed social network. The SLF vectors are learned first. Then, the SLF vectors for each node are concatenated as node features. Next, the node features are combined to form node pair features. The labels and features of the node pairs are then used to train the classification model for subsequent predictions.

$$\frac{\partial L_{(u)}}{\partial W_u^{out}} = -\sum\nolimits_{v\in N^n_{out}(u)\cup N^{ne}_{out}(u)}((1-F^-_{uv})W_v^{in}) + \sum\nolimits_{v\in N^p_{out}(u)\cup N^{non}_{out}(u)} F^-_{uv}W_v^{in} \quad (5)$$

$$\frac{\partial L_{(u)}}{\partial W_u^{in}} = -\sum\nolimits_{v\in N^n_{in}(u)\cup N^{ne}_{in}(u)}((1-F^-_{vu})W_v^{out}) + \sum\nolimits_{v\in N^p_{in}(u)\cup N^{non}_{in}(u)} F^-_{vu}W_v^{out} \quad (6)$$

Taking a single step in the coordinate ascent takes a linear amount of time, which means this approach is not particularly scalable to large networks. However, to reduce the training time, we replace $N^{non}_{out}(u)$ and $N^{non}_{in}(u)$ with sampled nodes from $N^{non}_{out}(u)$ and $N^{non}_{in}(u)$. There are two reasons for this: the first is that null relationships account for the vast majority of a signed social network, and therefore optimizing null relationships account for much of the training time. Second, many null relationships involve completely unrelated nodes, which add little useful information to the learning process.

Algorithm 1 summarizes the optimization procedure. When updating U_u^{out}, F^+_{uv} is computed $(d_u^{out}+n)$ times, where d_u^{out} is the outdegree of node u. Thus, the computational cost of updating U_u^{out} is $O((d_u^{out}+n)|k_+|)$. When updating W_u^{in}, F^-_{uv} is computed $(d_u^{in}+n)$ times, where d_u^{in} is the indegree of node u. Thus the computational cost of updating W_u^{in} is $O((d_u^{in}+n)|k_-|)$. Similarly, the computational costs of updating U_u^{in} and W_u^{out} are $O((d_u^{in}+n)|k_+|)$ and $O((d_u^{out}+n)|k_-|)$, respectively. In each iteration, the SLF vectors of $|V|$ nodes are updated, with a maximum T number of iterations. Since $d_u^{out}, d_u^{in} \ll |V|$ in signed social networks, and k_+, k_-, n, and T are constants, the overall computational cost of Algorithm 1 is $O(|V|)$.

5 EXPERIMENTS

In this section, we present the experiments used to evaluate the efficacy of the proposed SLF model on link prediction. We begin by introducing the networks, comparative methods, and evaluation metrics used. The experimental results follow with a parameter sensitivity analysis to conclude the section.

Algorithm 1: Optimization for learning SLF vectors

```
Input: the signed social network $G = (V, E^p, E^n, E^{ne})$,
       dimensions of positive latent factor space $k_1$,
       dimensions of negative latent factor space $k_2$, and
       maximum iterations $T$.
Output: signed latent factors $U^{out}$, $U^{in}$, $W^{out}$ and $W^{in}$.
Initialize $U_u^{out}$, $U_u^{in}$, $W_u^{out}$ and $W_u^{in}$ to random positive
 values less than 1.
repeat
    for u from 1 to $|V|$ do
        Update $U_u^{out}$ via Equation 3;
        Update $U_u^{in}$ via Equation 4;
        Update $W_u^{out}$ via Equation 5;
        Update $W_u^{in}$ via Equation 6;
        Set the negative elements of $U_u^{out}$, $U_u^{in}$, $W_u^{out}$ and
         $W_u^{in}$ to 0.
until converge;
```

5.1 Experimental Setup

5.1.1 Datasets. Four real-world signed social networks were used in these experiments: WikiElec, WikiRfa, Slashdot and Epinions. All four networks are collected from the SNAP repository [2], and all four are very sparse. The descriptive statistics for each of the four networks appear in Table 3.

- WikiElec and WikiRfa [5]: WikiElec is the voting network for the Wikipedia administrator elections, and WikiRfa is a

[2] http://snap.stanford.edu/data/index.html

Table 3: Network statistics.

Datasets	# nodes	# links	% positive links	% negative links	% neutral links
WikiElec	7,194	114,040	73.6	20.3	6.1
WikiRfa	10,885	137,966	73.0	20.8	6.2
Slashdot	82,140	549,202	77.4	22.6	0
Epinions	131,828	841,372	85.3	14.7	0

more recent version of WikiElec. These two datasets contain positive, negative, and neutral votes (links) with an average sparsity of 31.7 and 25.3, respectively. Notably, WikiElec and WikiRfa both contain just over 6% neutral votes - a sufficient proportion to test the impact of considering neutral links. Section 5.2 discusses how these neutral links have a positive effect on the prediction results.

- Slashdot is a friendship network for the technology-related news website [16]. It has an average sparsity of 13.4. Each user is allowed to tag another user as a friend (positive) or foe (negative).
- Epinions is a trust network for the consumer review site [16]. It has an average sparsity of 12.8. Its members can decide to trust (positive) or distrust (negative) other users based on the quality of their reviews.

5.1.2 Comparative Methods. Four state-of-the-art link prediction methods were selected as comparisons. The descriptions follow.

- Scalable embeddings for signed networks (SIGNet) [12]. This is a scalable feature learning framework suitable for signed networks. Its objective function aims to carefully model the social structures implicit in signed networks by reinforcing the principles of social balance theory.
- Matrix factorization (MF) [10]. Matrix factorization based model learns the low rank structures in a given network by decomposing the adjacency matrix of a signed social network into two low-rank matrices. Predicitons are then made by recovering the adjacency matrix.
- Link-oriented signed network embedding (LSNE) [7]. This method is an advancerd version of LINE [28]. It redefines the first order and second order proximities in LINE to suit signed social networks and learns the source and target embedding vectors by optimizing an objective function based on those proximities.
- Signed directed network embedding (SIDE) [14]. This is a Skip-Gram-based model that interprets negative edges as an indication of remoteness, and models asymmetric directions as biases. Structural balance theory combined with a random walk are used to generate multi-step connections for training.

As a self-comparison, we also tested SLF model without considering neutral links, which is denoted as SLF-degraded.

5.1.3 Evaluation Metrics. The specific task we assessed was to predict which unconnected node pairs would be connected and the signs of those links. We randomly selected 20% of the links in the networks and added them to the test set, and added the remaining links to the training set. Unconnected node pairs were undersampled to balance the distribution of social relationships, which meant null relationships accounted for around 75% of the training set and test set. Random selection was repeated 10 times independently; averaged results are reported.

Table 4: The impact of operators on link prediction with SLF.

		Avg	Con	L1	L2
AUC@p	WikiElec	0.927	**0.963**	0.785	0.786
	WikiRfa	0.935	**0.963**	0.790	0.800
	Slashdot	0.918	**0.936**	0.779	0.789
	Epinions	0.957	**0.962**	0.820	0.809
AUC@n	WikiElec	0.882	**0.941**	0.790	0.809
	WikiRfa	0.889	**0.942**	0.810	0.829
	Slashdot	0.930	**0.949**	0.854	0.877
	Epinions	0.912	**0.941**	0.886	0.880
AUC@non	WikiElec	0.930	**0.968**	0.795	0.805
	WikiRfa	0.940	**0.970**	0.809	0.824
	Slashdot	0.939	**0.955**	0.815	0.831
	Epinions	0.964	**0.967**	0.846	0.840
micro-F1	WikiElec	0.855	**0.901**	0.781	0.780
	WikiRfa	0.860	**0.898**	0.778	0.781
	Slashdot	0.869	**0.892**	0.798	0.803
	Epinions	0.913	**0.928**	0.829	0.830

Due to the imbalances common to signed social networks, we used the standard metrics area under curve (AUC) [9] and micro-F1 score [12] to evaluate the prediction performance. AUC and micro-F1 are blind to class distribution, and have been widely used to evaluate the quality of link prediction results. A larger AUC/micro-F1 value indicates better performance with an upper limit of 1.0, which represents a perfect prediction result. We used AUC to evaluate whether a method could distingguish one type of social relationships from the others. To this end, we considered one type of social relationships as the positive class and the remaining as the negative class. The prediction performance for positive, negative and null social relationships is denoted as AUC@p, AUC@n, and AUC@non, respectively. The second metric, micro-F1, reflects the overall performance in terms of whether a approach correctly labels the node pairs in the test set.

5.1.4 Parameter settings. For a fair comparison, we set the dimensions of the node representation to 64 for all comparative methods, but all other parameters were set according to the values recommended in their respective papers. The parameter settings for the SLF model were $k_1 = 32$, $k_2 = 32$, $n = 5$ and $p_0 = 0.001$.

Table 5: Comparison between SLF and other state-of-the-art methods on link prediction.

		SLF (proposed)	SIGNet	MF	LSNE	SIDE
AUC@p	WikiElec	**0.963**	0.909	0.933	0.750	0.823
	WikiRfa	**0.963**	0.784	0.922	0.719	0.802
	Slashdot	**0.936**	0.911	0.865	0.743	0.838
	Epinions	**0.962**	0.899	0.934	0.871	0.805
AUC@n	WikiElec	**0.941**	0.857	0.803	0.697	0.871
	WikiRfa	**0.942**	0.845	0.797	0.673	0.808
	Slashdot	**0.949**	0.854	0.773	0.715	0.884
	Epinions	**0.941**	0.905	0.895	0.865	0.832
AUC@non	WikiElec	**0.968**	0.888	0.910	0.752	0.894
	WikiRfa	**0.970**	0.769	0.901	0.722	0.898
	Slashdot	**0.955**	0.894	0.849	0.760	0.917
	Epinions	**0.967**	0.894	0.935	0.886	0.836
micro-F1	WikiElec	0.901	0.813	0.861	0.792	**0.952**
	WikiRfa	0.898	0.765	0.853	0.778	**0.954**
	Slashdot	0.892	0.822	0.835	0.798	**0.957**
	Epinions	**0.928**	0.846	0.903	0.862	0.914

5.1.5 Operators for node pair features. To determine which of the four operators should be used to combine the node features then construct node pair features, we assessed all four operators in terms of AUC@p, AUC@n, AUC@non and micro-F1. The results are shown in Table 4. The Con operator outperformed the other three operaters on all the four networks. This operator is suitable for all the comparison methods and, therefore, made the best choice as the feature operator.

5.2 Link Prediction

Table 5 shows the results of our link prediction experiments. From the results, we observe that:

- SLF was the most accurate method in all metrics, except for micro-F1 where SLF's performance was competitive. This result supports the efficacy of the proposed SLF model.
- SLF provided much better results than MF. MF is a traditional latent factor model that represents the opposite effects of a latent factor on social relationships according to the positive/negative values in the latent factor vectors. These results suggest that the signed latent factor vectors learned by the SLF model have the potential to greatly improve performance with link prediction tasks.
- SIGNet, LSNE, and SIDE learn the embedding vectors for each node by preserving structural approximates in the embedding space. SIDE was the only method to approach SLF's performance and had a better micro-F1 result. However, SIDE's overall performance was worse than SLF. These results suggest that signed latent factor vectors, which represent the relation between the nodes and the signed latent factors, may be more appropriate for link prediction tasks.

Table 6: The impact of neutral social relationships on link prediction with SLF.

		SLF	SLF-degraded
AUC@p	WikiElec	**0.963**	0.934
	WikiRfa	**0.963**	0.927
AUC@n	WikiElec	**0.941**	0.893
	WikiRfa	**0.942**	0.901
AUC@non	WikiElec	**0.968**	0.920
	WikiRfa	**0.970**	0.922
micro-F1	WikiElec	**0.901**	0.879
	WikiRfa	**0.898**	0.863

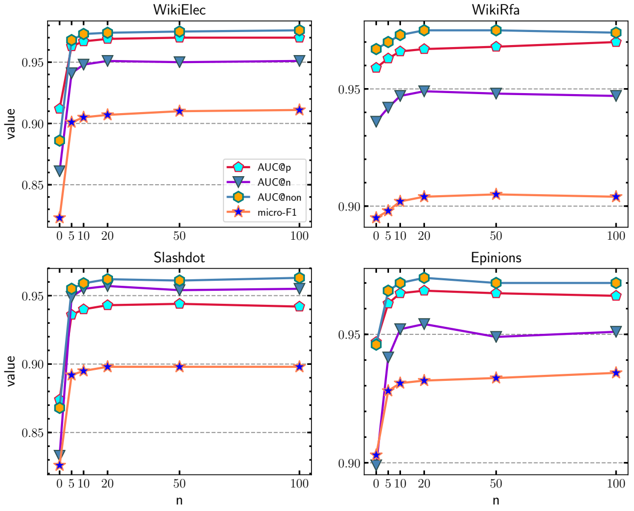

Figure 4: The impact of parameter n (sample size of null social relationships) on link prediction with SLF.

- SIGNet, MF, and LSNE showed better performance on Epinions than the other networks. Epinions has the largest proportion of positive links, which indicates that these methods may have difficulty predicting negative links in signed social networks.
- LSNE was not as competitive as SIGNet or SIDE, even though all three methods are based on learning the embeddings in signed social networks. However, LSNE learns the embeddings by preserving first- and second-order proximities. Hence, these results suggest this is not an appropriate technique for link prediction.

Table 6 reports the results of our comparison between SLF and SLF-degraded. Slashdot and Epinions do not contain any neutral links, so these experiments were limited to the WikiElec and WikiRfa networks. The results show that:

- SLF outperformed SLF-degraded in all four metrics, supporting our conjecture that considering neutral relationships improves link prediction performance.

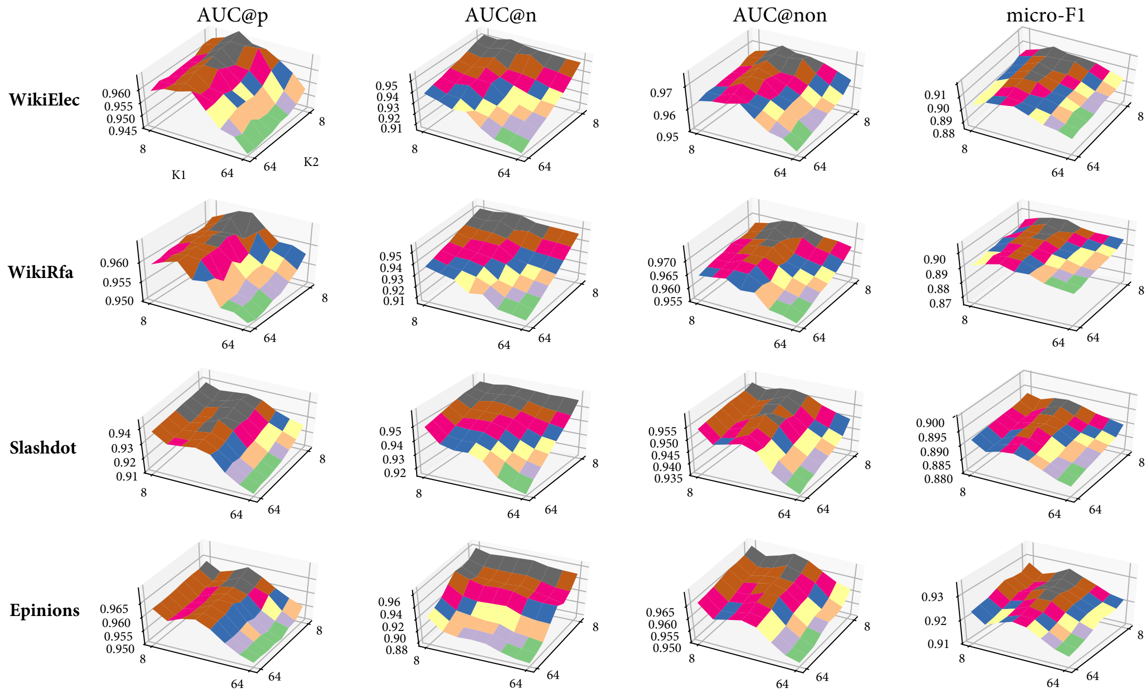

Figure 5: The impact of parameters k_1 and k_2 (demensions of the positive and negative SLF spaces) on link prediction with SLF.

- Further, the AUC@non result benefits most form the neutral relationships, which indicates that considering neutral relationships improves predictions about the propensity of a node to connect with other nodes.

5.3 Parameter Sensitivity

The most sensitive parameters in SLF are the sample size of the null relationships n, the initialization parameter for the logistic activation function p_0, and the dimensions of the SLF spaces k_1 and k_2. We analyzed each in turn as follows.

5.3.1 Impact of n. To investigate the effects of n (sample size of null social relationships), we fixed $k_1 = 32$, $k_2 = 32$ and $p_0 = 0.001$, then varied n in the set $\{0, 5, 10, 20, 50, 100\}$. The results are reported in Figure 4. From these results, we find:

- Ignoring null relationships reduces the performance of the SLF model, especially in terms of AUC@non, because too much information is lost.
- As n increases, performance initially increases but then stabilizes, making it feasible to use node pair sampling in the optimization.

5.3.2 Impact of p_0. To investigate the effects of p_0 (the effect of no feedback on the relationship), we fixed $k_1 = 32$, $k_2 = 32$ and $n = 5$, and varied p_0 in the set $\{0.001, 0.01, 0.1, 0.2\}$. The results are reported in Table 7. Here, we find:

- The smaller the p_0 (e.g., 0.001 and 0.01) the better SLF's performance, as indicated by the smoother variation in $f_a(x)$ when p_0 is small.

Table 7: The impact of parameter p_0 (the effect of no feedback on the relationship) on link prediction with SLF.

		p_0= 0.001	p_0= 0.01	p_0= 0.1	p_0= 0.2
AUC@p	WikiElec	0.963	0.957	0.935	0.937
	WikiRfa	0.963	0.955	0.928	0.930
	Slashdot	0.936	0.926	0.882	0.895
	Epinions	0.962	0.958	0.952	0.954
AUC@n	WikiElec	0.941	0.922	0.847	0.841
	WikiRfa	0.942	0.923	0.811	0.822
	Slashdot	0.949	0.936	0.798	0.784
	Epinions	0.941	0.918	0.895	0.900
AUC@non	WikiElec	0.968	0.960	0.922	0.922
	WikiRfa	0.970	0.960	0.907	0.912
	Slashdot	0.955	0.943	0.886	0.902
	Epinions	0.967	0.962	0.953	0.955
micro-F1	WikiElec	0.901	0.894	0.869	0.861
	WikiRfa	0.898	0.889	0.857	0.850
	Slashdot	0.892	0.887	0.859	0.856
	Epinions	0.928	0.923	0.912	0.911

- However, with a large p_0, the minimum values for F_{uv}^+ and F_{uv}^- are also large, which makes it more difficult to distinguish null relationships from the others. Hence, performance declines.

5.3.3 *Impact of k_1 and k_2.*

To investigate the effects of k_1 and k_2 (demensions of the positive and negative SLF spaces), we fixed $n = 5$ and $p_0 = 0.001$, and varied both k_1 and k_2 in the set $\{8, 16, 24, 32, 40, 48, 56, 64\}$. The results are reported in Figure 5, where we find:

- As k_1 increases, SLF's overall performance increases initially but then decreases. When k_1 is small, SLFs does not have sufficient representation capacity but, when k_1 is large, SLF has a tendency to overfit.
- Compared to k_2, k_1 has a greater impact on performance. The reason for this is because signed social networks usually have more positive links than negative ones, and the k_1 is more closely related to the positive links.

6 CONCLUSION

This paper outlines a signed latent factor model, called SLF, for solving link prediction problems in signed social networks. We designed two types of signed latent factors. The merit of the proposed model is that SLF considers four different types of social relationships - positive, negative, neutral and no relationship at all (i.e., null). Each type of relationship is linked to the comprehensive effects of positive and negative signed latent factors. The SLF model is based on a 3-step procedure that uses signed latent factor vectors to calculate four scores. Each score corresponds to the four types of social relationships based on their sociological meaning. Further, we introduced a nonlinear factor to reduce the potential for overfitting. Most methods ignore neutral links. However, through self-comparisons of SLF, we find that considering neutral relationships benefits performance. Further experiments with several state-of-the-art methods on four real-world signed social networks demonstrate the advancements the SLF model makes to link prediction problems in signed social networks.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (Nos. 61572369, 61822113, and 61702355), in part by the Key Projects of Guangdong Natural Science Foundation (No. 2018B030311003), in part by the MQNS (No. 9201701203), the MQEPS (Nos. 90275252 and 96804590), the MQRSG (No. 95109718), and in part by the Investigative Analytics Collaborative Research Project between Macquarie University and Data61 CSIRO.

REFERENCES

[1] Nahla Mohamed Ahmed and Ling Chen. 2016. An efficient algorithm for link prediction in temporal uncertain social networks. *Information Sciences* 331 (2016), 120–136.

[2] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. 2008. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* 9, Sep (2008), 1981–2014.

[3] Albert-Laszlo Barabâsi, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. 2002. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications* 311, 3-4 (2002), 590–614.

[4] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2014. Who to follow and why:link prediction with explanations. In *Acm Sigkdd International Conference on Knowledge Discovery & Data Mining*.

[5] Gerard Cabunducan, Ralph Castillo, and John Boaz Lee. 2011. Voting Behavior Analysis in the Election of Wikipedia Admins. In *International Conference on Advances in Social Networks Analysis & Mining*.

[6] Kai Yang Chiang, Nagarajan Natarajan, Ambuj Tewari, and Inderjit S. Dhillon. 2011. Exploiting longer cycles for link prediction in signed networks. In *Acm International Conference on Information & Knowledge Management*.

[7] Dongfang Du, Hao Wang, Tong Xu, Yanan Lu, Qi Liu, and Enhong Chen. 2017. Solving link-oriented tasks in signed network via an embedding approach. In *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*. IEEE, 75–80.

[8] Daniel M Dunlavy, Tamara G Kolda, and Evrim Acar. 2011. Temporal link prediction using matrix and tensor factorizations. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5, 2 (2011), 10.

[9] James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36.

[10] Cho Jui Hsieh, Kai Yang Chiang, and Inderjit S. Dhillon. 2012. Low rank modeling of signed networks. In *Acm Sigkdd International Conference on Knowledge Discovery & Data Mining*.

[11] Cho-Jui Hsieh and Inderjit S Dhillon. 2011. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1064–1072.

[12] Mohammad Raihanul Islam, B Aditya Prakash, and Naren Ramakrishnan. 2018. Signet: Scalable embeddings for signed networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 157–169.

[13] Rodolphe Jenatton, Nicolas L Roux, Antoine Bordes, and Guillaume R Obozinski. 2012. A latent factor model for highly multi-relational data. In *Advances in Neural Information Processing Systems*. 3167–3175.

[14] Junghwan Kim, Haekyu Park, Ji-Eun Lee, and U Kang. 2018. Side: representation learning in signed directed networks. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 509–518.

[15] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.

[16] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Predicting Positive and Negative Links in Online Social Networks. In *International Conference on World Wide Web*.

[17] Wu-Jun Li, Dit-Yan Yeung, and Zhihua Zhang. 2011. Generalized latent factor models for social network analysis. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, Spain*. 1705.

[18] Xiaoming Li, Hui Fang, and Jie Zhang. 2017. Rethinking the Link Prediction Problem in Signed Social Networks.. In *AAAI*. 4955–4956.

[19] Xiaoming Li, Hui Fang, and Jie Zhang. 2018. FILE: A Novel Framework for Predicting Social Status in Signed Networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[20] Chih-Jen Lin. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural computation* 19, 10 (2007), 2756–2779.

[21] Julian Mcauley and Jure Leskovec. 2012. Learning to discover social circles in ego networks. In *International Conference on Neural Information Processing Systems*. 539–547.

[22] Aditya Krishna Menon and Charles Elkan. 2010. A log-linear model with latent features for dyadic prediction. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 364–373.

[23] Aditya Krishna Menon and Charles Elkan. 2011. Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases*. Springer, 437–452.

[24] Mark EJ Newman. 2001. Clustering and preferential attachment in growing networks. *Physical review E* 64, 2 (2001), 025102.

[25] Thin Nguyen, Dinh Q Phung, Brett Adams, and Svetha Venkatesh. 2011. Towards Discovery of Influence and Personality Traits through Social Link Prediction.. In *ICWSM*. 566–569.

[26] Dongjin Song and David A Meyer. 2015. Recommending Positive Links in Signed Social Networks by Optimizing a Generalized AUC.. In *AAAI*. 290–296.

[27] Jiliang Tang, Charu Aggarwal, and Huan Liu. 2016. Recommendations in Signed Social Networks. In *International Conference on World Wide Web*.

[28] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1067–1077.

[29] Jiliang Tang, Chang Yi, Charu Aggarwal, and Huan Liu. 2015. A Survey of Signed Network Mining in Social Media. (2015).

[30] Jihang Ye, Cheng Hong, Zhu Zhe, and Minghua Chen. 2013. Predicting positive and negative links in signed social networks by transfer learning. In *International Conference on World Wide Web*.

[31] Linhong Zhu, Dong Guo, Junming Yin, Greg Ver Steeg, and Aram Galstyan. 2016. Scalable temporal latent space inference for link prediction in dynamic social networks. *IEEE Transactions on Knowledge and Data Engineering* 28, 10 (2016), 2765–2777.