The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

FILE: A Novel Framework for Predicting Social Status in Signed Networks

Xiaoming Li, Hui Fang,* Jie Zhang
School of Computer Science and Engineering, Nanyang Technological University, Singapore
*Research Institute for Interdisciplinary Sciences and School of Information Management and Engineering
*Shanghai University of Finance and Economics, China
lixiaoming@ntu.edu.sg, fang.hui@mail.shufe.edu.cn, zhangj@ntu.edu.sg

Abstract

Link prediction in signed social networks is challenging because of the existence and imbalance of the three kinds of social status (positive, negative and no-relation). Furthermore, there are a variety types of no-relation status in reality, e.g., strangers and frenemies, which cannot be well distinguished from the other linked status by existing approaches. In this paper, we propose a novel Framework of Integrating both Latent and Explicit features (FILE), to better deal with the no-relation status and improve the overall link prediction performance in signed networks. In particular, we design two latent features from latent space and two explicit features by extending social theories, and learn these features for each user via matrix factorization with a specially designed ranking-oriented loss function. Experimental results demonstrate the superior of our approach over state-of-the-art methods.

Introduction

Signed social networks have been widely adopted by online communities over the last few years, as they better reflect real-life human relationships than unsigned networks (Tang et al. 2016). Under the structure, three kinds of social status exist between two users: positive (trust or friend), negative (distrust or foe), and no-relation. For example, in the Wikipedia vote network, a user may vote a candidate entity as positive or negative, and can also choose not to vote and thus maintain no-relation with the entity. The increasing interest in signed networks has heightened the need to rethink the link prediction problem (Liben-Nowell and Kleinberg 2007), since it becomes more challenging in the scenario of signed networks than unsigned networks which consider only two kind of social status (linked or not).

Most studies (Leskovec, Huttenlocher, and Kleinberg 2010a; Chiang et al. 2011; Hsieh, Chiang, and Dhillon 2012; Ye et al. 2013) on link prediction in signed networks focus on predicting the sign of a link, i.e., assigning either a positive or negative sign to any pair of users. They show that positive links and negative links can be distinguished with a high accuracy. However, these studies simply assume that it is already known whether there is a link between any two users, which is invalid in real-world scenarios. Recently, a few methods have been proposed by also considering no-relation as a meaningful social status to facilitate link prediction. For example, Song et al. (2015) demonstrate that leveraging the no-relation status can improve the prediction of positive links. However, they only focus on the prediction performance of positive links but cannot predict the no-relation status. Li et al. (2017b) carry out the first attempt to extend the link prediction to a more realistic setting by also predicting the no-relation status. They show that no-relation can be distinguished from positive and negative links, through a feature-based model, where the features are extracted from social theories. However, this model is limited to the assumption that users have the same criteria to link with others (Nguyen et al. 2011), which is unrealistic. For example, some users might be more willingly to connect to others while some are more influential and easily connected by others (Nguyen et al. 2011).

In fact, the link prediction problem in signed social networks becomes rather difficult mainly due to the diversity of no-relation. It is conceivable that most pairs of users with no-relation have limited common connections (*Stranger*). However, in reality, many user pairs keep the no-relation status even though they have many common connections (*Frenemy*). For example, in the Epinions dataset[1], $40,779$ out of $94,732$ user pairs who share more than 100 common neighbors still have no-relation with each other. It is very easy to mispredict those users, who have many common neighbors but are not linked, with a linked status.

In this paper, we propose a novel Framework of Integrating both Latent and Explicit features (FILE), to better deal with the no-relation status in signed networks. The key idea is to design two essential parts to represent the link formation probability. The first part is the social linkage criteria from the perspective of individual users, and the second part is the external social influence from the perspective of user pairs. Specifically, we design two latent features for the first part. One is the propensity to connect to others, namely the activeness, and the other is the propensity to be connected by others, namely the popularity. We train these two features via the matrix factorization technique with a ranking-oriented loss function, and then we represent the linkage likelihood as the inner product between the corresponding

[1] www.trustlet.org/epinions.html

two user vectors. For the second part, we design the explicit features extracted from social theories (e.g., balance theory and status theory) to represent the external social influence. Both parts are indispensable, since the lack of the latent features will lead to the misprediction between a frenemy and a friend, while the model without explicit features will mispredict two strangers as a linked one. The extensive experiments on four real-world datasets demonstrate the effectiveness of our framework on link prediction in signed networks.

All in all, the contributions of this work are as follows:

- We propose a novel link prediction framework which integrates social explicit features into a latent model. We demonstrate that this can significantly improve the prediction of positive link, negative link and no-relation.
- We take a deep investigation on the no-relation status. We empirically show that two types of no-relation status widely exist in real-world datasets, and the proposed framework can well handle both of the two types.

Related Work

Link prediction in unsigned networks has been well studied during the past decade. It mainly calculates a "link formation score" for two users to indicate their probability to be linked in near future (Liben-Nowell and Kleinberg 2007). Popular calculation metrics include: the number of common neighbors, Adamic/Adar Index (Adamic and Adar 2003), Jaccard Coefficient (Newman 2001), and Resource Allocation Index (Zhou, Lü, and Zhang 2009). These metrics are derived from the neighborhood structure. Meanwhile, the features related to the path between two users are also used to compute the similarities of the user pair, like Katz (Liben-Nowell and Kleinberg 2007), Vertex Collocation Profile (Lichtenwalter and Chawla 2012) and ProfFlow (Lichtenwalter, Lussier, and Chawla 2010). Popular supervised methods include: feature-based classification models (Al Hasan et al. 2006) and latent feature models (Menon and Elkan 2011). However, link prediction in unsigned networks considers only two possible connection status of two users, i.e., linked or not-linked, while three types of social status exist in signed networks.

In signed networks, Leskovec et al. (2010a) adopt a regression model with triangle-based features to predict the sign (i.e., positive or negative) between each two users. Besides, k-cycle-based features are proposed in (Chiang et al. 2011) where triangle-based features ($k = 3$) are specially explored. It also shows that longer cycles ($k = 5$) significantly benefit sign prediction, while the performance gain is not significant beyond $k = 5$. Papaoikonomou et al. (2014) leverage the pattern of frequent subgraph among user pairs to predict link status. These methods can well distinguish positive and negative links, however, they are all based on an unrealistic assumption where it is already known whether there is a link between any two users.

Hsieh et al. (2012) state that three social status exist in signed networks, which are positive, negative and no-relation. They treat no-relation as a potential status to be linked, and propose a matrix factorization model to infer the signs of those "potential links" which currently are no-relation. However, they ignore that no-relation could be stable and possibly will not transform to a linked one. Song and Meyer (2015) adopt a low-rank model to recommend positive links, which learns latent features by capturing the intuition that linked pairs have a different status with no-relation, and no-relation status can help to better embed users. Kumar et al. (2016) adopt a recursive model for link prediction in weighted signed networks, where no-relation can be treated as a special case in which the link value is zero. However, this model still cannot predict no-relation since it only predicts the link status with non-zero value. Li et al. (2017a) are the first to treat no-relation as a future status for link prediction. They derive topological features based on six social theories, and adopt a simple regression model to distinguish these three status. They show that no-relation status can be distinguished from positive and negative links by social theory-based features. However, they ignore individual differences on the social linkage criteria.

Preliminaries

Problem Formulation

We formally define the problem as: given a signed social network $S \in \mathbb{R}^{n \times n}$ (n is the number of users in the network) with $S_{ij} \in \{1, 0, -1\}$, we aim to rank all the user pairs (i, j) with $S_{ij} = 0$ in the present, by the probability of transforming to positive links, negative links, or maintaining no-relation in the future. We argue that our problem setting is more comprehensive and realistic compared to the previous studies. Rather than classifying a user pair as a specific social relation, we adopt a ranking mechanism and try to answer a more practical question: "Of user pairs (i, j) and (i, k), which pair is more likely to become friends (or enemies)?" The obtained ranking list can be directly utilized in real-world applications like social recommendation.

Data Analysis

Previous analysis on data patterns in signed networks (Leskovec, Huttenlocher, and Kleinberg 2010b; Tang et al. 2016) are preliminary and focus only on the comparisons between positive and negative links. We now re-investigate data patterns by also considering the no-relation status. Our analysis is performed on four real-world signed networks: Epinions, Slashdot[2], Wikipdia RFA[3] and Bitcoins[4].

Data imbalance. From Table 1, we can see that no-relation accounts for the majority of social status, and the number of no-relation is much larger than linked ones. Meanwhile, the proportions of those social relations vary in different datasets, requiring the robustness of the proposed method on various scenarios. It should be noted that as ranking metrics (e.g., AUC) are relatively effective for evaluating and distinguishing machine learning techniques in imbalanced scenarios, we also use ranking metrics instead of ac-

[2] snap.stanford.edu/data/soc-Slashdot0902.html

[3] snap.stanford.edu/data/wiki-RfA.html

[4] cs.umd.edu/~srijan/wsn

Table 1: Dataset statistics.

	Epinions	Slashdot	Wikipedia	Bitcoin
Users	131,828	82,140	9,654	3,783
Positive links (P)	717,667	425,072	87,766	22,650
Negative links (N)	123,705	124,130	16,788	1,536
No-relation (U)	1.73×10^{10}	6.7×10^{9}	9.3×10^{7}	1.4×10^{7}
U with CN=0	1.72×10^{10}	6.6×10^{9}	8.5×10^{7}	1.3×10^{7}
U with $1 \leq$CN≤ 50	1.6×10^{9}	9.7×10^{7}	7.2×10^{6}	1.1×10^{6}
U with CN$>$50	234,793	9,752	3,390	13

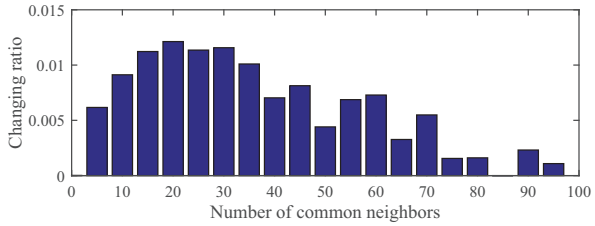


Figure 1: The distribution of no-relation changing ratio.

curacy metrics for reasonable evaluations and comparisons of different approaches in our experiments.

Stranger v.s. Frenemy. The statistics in Table 1 demonstrate the existence of two kinds of no-relation status as there are a substantial number of no-relation pairs with few common neighbors (CN=0) or many common neighbors (CN>50). For example, in Epinions dataset, the number of common neighbors for no-relation user pairs ranges from 1 to 2,059. In other words, even a user pair with 2,059 common neighbors may still have no link with each other.

We further check whether these no-relation pairs are stable over time in Epinions dataset as it contains the information about timestamp of every link formation over 30 months. Figure 1 shows the changing ratio of the no-relation user pairs after 15 months. Y axis is computed as the number of no-relation user pairs with a certain number of common neighbors who are linked after 15 months divided by the number of no-relation user pairs with the certain number of common neighbors in the present. We observe that no-relation status of user pairs can be stable over time even though they have many common neighbors, and user pairs with more common neighbors may not have a higher probability of being linked in the future. On the contrary, when the number of common neighbors is larger than 20, no-relation status becomes more stable with more common neighbors.

As we have known, the core task of link prediction is to calculate a "link formation score" for a user pair. Since both user pairs of frenemies and strangers belong to the same no-relation class, they are expected to have a similar score. However, most existing approaches relying on network topological features cannot achieve this simple goal as frenemies and strangers have quite different topological features (e.g., the number of common neighbors). Therefore, we clarify that the core task of link prediction in signed networks is more suitable to be explicitly defined as "how to design a link score function to generate similar scores for

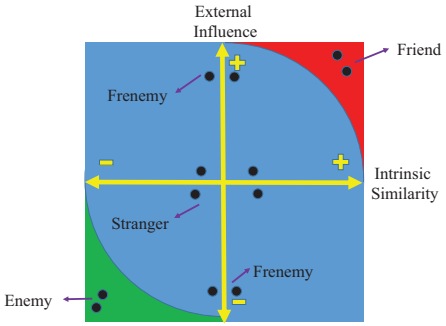


Figure 2: Illustration of the influential social components.

frenemies and strangers, meanwhile to be able to distinguish them from positive and negative links".

In view of psychosocial theories, both intrinsic personality (Duck and Craig 1978) and external influence from mutual neighbors (Bukowski and Hoza 1989) affect social relationship formation. Thus, it is reasonable to explain the stranger relationship as a lack of external influence, and the frenemy relationship as a lack of intrinsic personality similarity. Figure 2 illustrates the differences among these social relations. Inspired by this, in our FILE framework, we derive a user's latent features in the latent space to represent the intrinsic personality, and design explicit features based on network topology to represent the external influence.

The FILE Framework

In this section, we describe the FILE framework incorporating both latent and explicit features for link prediction in signed networks. We first present the two types of features in detail, and then introduce our designs of the link score function and the optimization method.

Latent Features

A signed network can be represented by a signed adjacency matrix S ($S \in \mathbb{R}^{n \times n}$) associated with the n users and links in the network, where $S_{ij} = 1$ indicates a positive link from user i to j, $S_{ij} = -1$ a negative link from i to j, and $S_{ij} = 0$ no-relation from i to j representing the majority of the entry values in S. Since this kind of matrices of signed networks has the low-rank property (Hsieh, Chiang, and Dhillon 2012), matrix factorization technique can be deployed to learn users' latent features. Specifically, S can be decomposed into two low-rank matrices U and V, where $U^T V \approx S$ ($U, V \in \mathbb{R}^{n \times r}, r \ll n$). We call both $u_i \in U$ and $v_i \in V$ as user i's latent vectors, being referred to as the activeness and popularity respectively. For a certain user pair i and j, the probability of link formation simultaneously depends on both u_i and v_j, i.e., whether i is active and has more tendency to "trust" (or distrust) others, and whether j is popular and more probably to be trusted (or distrusted) by others. A higher value of $u_i^T v_j$ indicates a higher probability to form a positive link. Conversely, a lower value of $u_i^T v_j$ implies a higher probability to form a negative link.

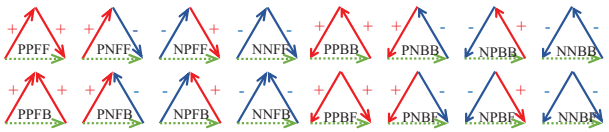


Figure 3: Illustration of 16 types of triads.

To sum up, given a pair of users (i, j), and the r-dimensional features u_i (activeness of user i) and v_j (popularity of user j), we define the link formation probability from user i to j as:

$$\mathcal{L}^1(i,j) = u_i^T \cdot v_j \tag{1}$$

Explicit Features

Explicit features capture social influences from the surrounding neighborhoods around a user pair, and can be formulated from the network topology. We claim that any valuable and reasonable features identified in the literature can be incorporated into the FILE framework (Leskovec, Huttenlocher, and Kleinberg 2010b; Li, Fang, and Zhang 2017b) as they contribute new information to social influences. In our framework, to show the effectiveness of the explicit features part, we design two explicit features by extending the balance theory and status theory. According to the two social theories, each common neighbor will bring either a positive or a negative influence. As shown in Figure 3, there are in total 16 types of triads formed by a pair of users and their mutual neighbor (p and n denote the positive and negative signs, and f and b represent the link directions of forward and backward respectively).

As indicated in the balance theory, each mutual friend brings a positive influence which makes two users more likely to generate a positive link, while a neighbor incurs a negative influence if she is one's friend but the other's enemy. Therefore, we check whether the positive or negative influence is dominant in the balance theory via:

$$f^1 = (|\text{ppff}|+|\text{ppfb}|+|\text{ppbf}|+|\text{ppbb}|+|\text{nnff}|+|\text{nnfb}|+|\text{nnbf}|+|\text{nnbb}|) \\ -(|\text{pnff}|+|\text{pnfb}|+|\text{pnbf}|+|\text{pnbb}|+|\text{npff}|+|\text{npfb}|+|\text{npbf}|+|\text{npbb}|)$$

where $|\cdot|$ represents the number of respective type of triads.

In the status theory, a neighbor can imply the status difference between a user pair. For example, for **ppff**, given a user pair (i, j) and their neighbor w, the links $i \to w$ and $w \to j$ are both positive. Based on the status theory, it suggests that j's status is higher than w while w's status is higher than i. Therefore, the link $i \to j$ is more likely to be positive since the status of j is higher than i. We thus quantify the overall influence in the status theory via:

$$f^2 = |\text{ppff}|+|\text{nnbb}|+|\text{pnfb}|+|\text{npbb}|-(|\text{nnff}|+|\text{ppbb}|+|\text{npfb}|+|\text{pnfb}|)$$

For these two features, a higher positive (negative) value indicates a higher probability to form a positive (negative) link. A value close to 0 suggests they will be more likely to keep no-relation. We conduct the One-Way ANOVA test on explicit features to evaluate their effectiveness, and both the two features pass the test at the significance level of 0.01 (p-value < 0.01), suggesting that they can reasonably distinguish the three kinds of social status.

Note that we do not aim to come up with an exhaustive list of explicit features in this work. A more comprehensive list of explicit features can be found in (Leskovec, Huttenlocher, and Kleinberg 2010b; Li, Fang, and Zhang 2017b). Our experimental results show that with only the above two explicit features, our approach can already achieve better results than other existing approaches.

Link Score Function

The link score function is defined as follows:

$$\mathcal{L}(i,j) = \overbrace{N(u_i^T \cdot v_j)}^{\text{Latent}} + \overbrace{\sum_k \alpha_w * N(f_{ij}^w)}^{\text{Explicit}} \tag{2}$$

As aforementioned, both latent and explicit features are indispensable since the lack of any will lead to the misprediction of no-relation. In view of this, we first define a threshold rule for the link formation: there will be a positive link if the link score is larger than 1, and a negative link if the link score is smaller than -1. We bound the value of each part (Latent or Explicit) by $(-1, 1)$, which indirectly constrains that only the combination of two parts can successfully induce an either positive or negative link.

In Equation 2, u_i is user i's latent feature of activeness, v_j is user j's latent feature of popularity, f_{ij}^w $(w \in \{1, 2\})$[5] is an explicit feature for user pair $\{i, j\}$, α_w is the corresponding weight with $\sum_w \alpha_w = 1$, $N(\cdot)$ is the function which normalizes the corresponding values of features into $(-1, 1)$. Hence, the link score function is bounded and $L_{ij} \in (-2, 2)$. Based on the previous analysis, if L_{ij} is within $(-1, 1)$, there will be no link from i to j. If $L_{ij} \geq 1$, there will be a positive link from i to j, and if $L_{ij} \leq -1$, there will be a negative link from i to j.

Normalization Function. It normalizes the feature values into range $(-1, 1)$. Here, we formulate it as follows.

$$N(x|\theta) = \frac{1 - \exp(-\theta x)}{1 + \exp(-\theta x)} \tag{3}$$

The sigmoid distribution well captures the property of link formation that the value increases at a lower speed when i and j already show a high probability to establish a link. The selection of θ mainly depends on the scale of the corresponding feature. In this work, we normalize the two explicit features by making them to be scaled within the same order of magnitude. To this end, we set θ as the reciprocal of the median value of the corresponding feature.

Optimization

The traditional square loss is not suitable for our problem, because instead of caring about the absolute prediction error, we focus on the ranking performance. That is to say, for example, given a possibly positive link $S_{ij} = 1$, there should not incur any loss if predicted $L_{ij} \geq 1$. Therefore, in view of Equation 2, the loss function is defined as:

$$\min \sum_{S_{ij}=1} I(L_{ij} \geq 1) + \sum_{S_{ij}=0} I(|L_{ij}| < 1) + \sum_{S_{ij}=-1} I(L_{ij} \leq -1) \tag{4}$$

[5]Note that as indicated in the explicit features part, more explicit features can be designed and incorporated into Equation 2.

where $I(\cdot)$ is the $0/1$ indicator function that if the condition in $(\cdot)$ comes true, we get 0 loss, otherwise 1 loss. We aim to find a surrogate function to replace $I(\cdot)$ because it is non-convex. Considering our link score function in Equation 2, the ultimate goal of the objective function can be interpreted as to make L_{ij} as large as possible if $S_{ij} = 1$, meanwhile make L_{ij} as small as possible if $S_{ij} = -1$. As for $S_{ij} = 0$, we make $|L_{ij}|$ to be closer to 0. In view of this rationale, we design the objective function as follows:

$$\min \sum_{S_{ij}=1} (1 - L_{ij}) + \sum_{S_{ij}=0} (L_{ij}^2 - 1) + \sum_{S_{ij}=-1} (L_{ij} + 1) \quad (5)$$

To construct the equivalent reduced form for Equation 5 and add regularizers to avoid overfitting, the loss function F can be rewritten as follows:

$$\min_{U,V} \frac{1}{2} \sum_i \sum_j (1 - S^2)L^2 - SL + \frac{\lambda_1}{2}\|U\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2 \quad (6)$$

We then adopt stochastic gradient descent (SGD) to learn the values of parameters and variables. In particular, we first make $x = 1/(1+e^{-u_i^T v_j})$, $\Delta_1 = 2x + \sum_w \alpha N(f_{ij}^w) - 1$, and $\Delta_2 = 2x(1-x)$. Then the corresponding partial derivatives are computed as follows:

$$\frac{\partial F}{\partial u_i} = \sum_j \left((1 - S^2)\Delta_1\Delta_2 - S\Delta_2\right) * v_j + \lambda_1 u_i \quad (7)$$

$$\frac{\partial F}{\partial v_j} = \sum_i \left((1 - S^2)\Delta_1\Delta_2 - S\Delta_2\right) * u_i + \lambda_2 v_j \quad (8)$$

Algorithm 1 Optimization process

Input: Matrix S, learning rate β, iteration time T, and converge threshold c
Initialize: $t = 0$, calculate $f_{ij} \in E$, generate U_0, V_0
repeat
  $t = t + 1$;
  $U_{t+1} = U_t - \beta \frac{\partial F}{\partial U_t}$ based on Equation. 7;
  $V_{t+1} = V_t - \beta \frac{\partial F}{\partial V_t}$ based on Equation. 8;
until Converge
Output: U, V

Algorithm 1 summarizes the optimization procedure of the SGD. The time complexity of the algorithm is $O(trn)$, where t is the number of iterations, r is the number of latent features, n is the number of observations in the network.

Experiments

We conduct experiments on four real-world datasets, and compare our approach with five state-of-the-art approaches in terms of ranking metrics.

Experimental Setting

As shown in Table 1, four datasets are used in the experiments, which are Epinions, Slashdot, Wikipedia RFA and Bitcoin. To make a more comprehensive evaluation, we directly generate three datasets from each dataset, and each new generated dataset shows unique distribution of $|P|:|U|:|N|$, where $|P|$, $|U|$, $|N|$ are the numbers of positive links, no-relation, and negative links respectively. Specifically, we sample 10% data for each of the three large datasets (Epinions, Slashdot, Wikipedia) and select the data entries filtered by user degree d (≥ 10, ≥ 25, ≥ 50). The benefits of this setting include: 1) in the real-world offline case, people keep 40 friends on average (Express.co.uk 2017) and an online user has about 338 friends on average (Mazie 2016). Therefore, it is more realistic to check users with a high degree. This sampling strategy is widely adopted in the previous studies (Liben-Nowell and Kleinberg 2007); 2) we can test the model robustness under different scenarios in terms of data sparsity and size. The statistics of the datasets are summarized in Table 2 where we use 'name@degree' to represent a specific dataset, e.g., Epinions@10 (or E@10) is the dataset about Epinions with $d \geq 10$.

Table 2: 12 datasets used in the experiments.

Datasets	Positive	No-relation	Negative	Ratio
Epinions@10	38,452	4,017,624	8,180	5:491:1
Epinions@25	26,732	797,001	4,367	6:182:1
Epinions@50	17,039	233,624	2,346	7:99:1
Slashdot@10	22,551	1,544,792	2,666	8:579:1
Slashdot@25	16,097	359,568	1,331	12:270:1
Slashdot@50	11,023	119,265	756	14:157:1
Wikipedia@10	2,585	172,644	332	7:520:1
Wikipedia@25	363	12,594	39	9:322:1
Wikipedia@50	131	3,454	15	8:230:1
Bitcoin@10	10,863	361,590	868	12:461:1
Bitcoin@25	5,093	43,780	411	12:106:1
Bitcoin@50	2,048	7,551	202	10:37:1

Evaluation Metrics. We use the standard 5-fold cross-validation for training and testing, and utilize GAUC (Generalized AUC over $+1$, 0 and -1) (Song and Meyer 2015) to measure the overall ranking performance, formulated as:

$$\frac{1}{|P|+|N|}\left(\frac{1}{|U|+|N|}\sum_{a_i\in P}\sum_{a_s\in U\cup N} I\left(L(a_i) > L(a_s)\right) + \frac{1}{|U|+|P|}\sum_{a_j\in N}\sum_{a_t\in U\cup P} I\left(L(a_j) < L(a_t)\right)\right)$$

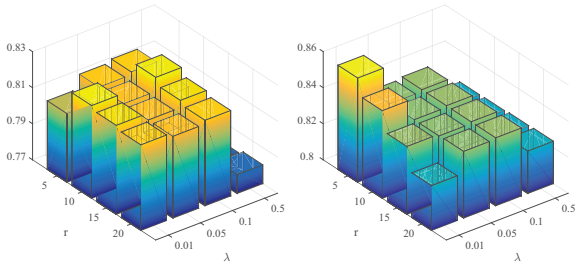
where $L(\cdot)$ is the link score function. GAUC is an extension of AUC, and provides a ranking metric considering the three kinds of link status.

The other metric is precision@top k. In signed networks, we have both positive and negative precision@top k, which are defined as the ratio of positive (or negative) links in the top (or bottom) k predictions, respectively. These two metrics assess the performance of link recommendation, as the top k list is more crucial for applications like recommendation systems, whereas the negative top k is useful for security-related applications.

Benchmarking Approaches. We conduct comparisons with five state-of-the-art approaches, including feature-based models: Common Neighbors (CN) (Liben-Nowell

Table 3: Performance of different methods. The best performance is highlighted in bold, and the second-best one is marked by *. 'Improvement' indicates the improvement of FILE over the model having the highest performance other than FILE.

Datasets	CN	LRM	BPRMF	OptGAUC	SFM	FILE	Improvement
Epinions@10	0.557	0.719	0.743	0.764*	0.738	**0.826**	8.12%
Epinions@25	0.563	0.731	0.730	**0.843**	0.742	0.842*	−0.19%
Epinions@50	0.557	0.741	0.696	0.789*	0.784	**0.823**	4.31%
Slashdot@10	0.525	0.697	0.658	0.721*	0.708	**0.823**	14.15%
Slashdot@25	0.520	0.747	0.639	0.792*	0.757	**0.838**	5.81%
Slashdot@50	0.502	0.760	0.685	0.827*	0.771	**0.856**	3.51%
Wikipedia@10	0.509	0.534	0.561	0.652	0.665*	**0.729**	9.62%
Wikipedia@25	0.593	0.508	0.577	0.714*	0.605	**0.727**	1.82%
Wikipedia@50	0.540	0.551	0.568	0.625*	**0.643**	0.595	−8.07%
Bitcoin@10	0.512	0.627	0.607	0.683*	0.682	**0.717**	4.98%
Bitcoin@25	0.555	0.706	0.609	0.715	0.716*	**0.723**	0.98%
Bitcoin@50	0.557	0.711	0.665	0.692	0.710*	**0.716**	0.85%



(a) Epinions@50 (b) Slashdot@50

Figure 4: Performance as parameters change w.r.t. GAUC.

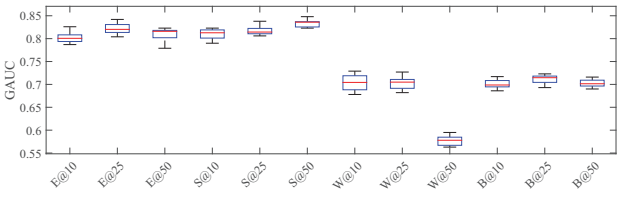


Figure 5: Performance fluctuations across datasets with different parameter combinations.

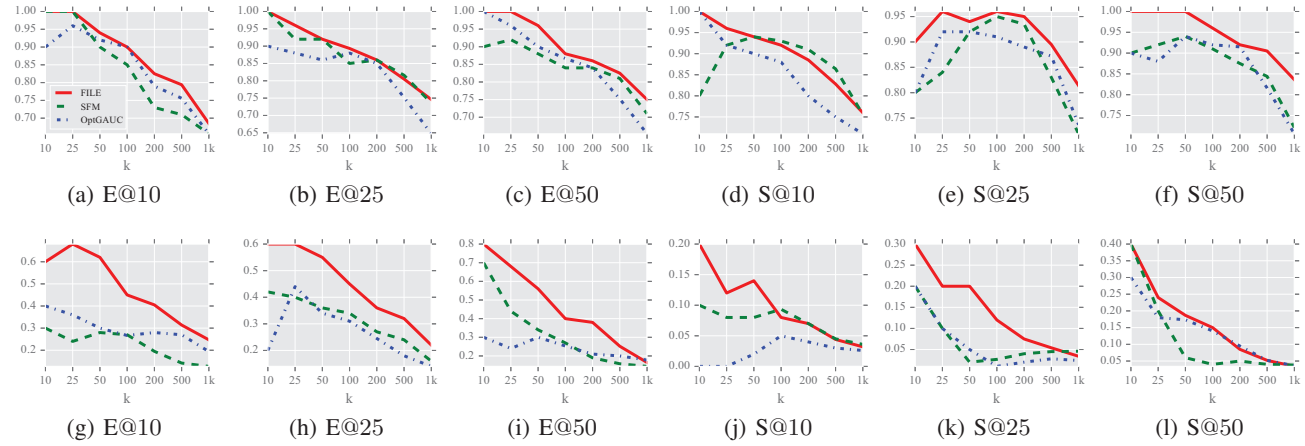
and Kleinberg 2007) and Social Feature Model (SFM) (Li, Fang, and Zhang 2017a); latent models: Low Rank Modeling (LRM) (Hsieh, Chiang, and Dhillon 2012) and ranking based latent models of Bayesian Personalized Ranking (BPRMF) (Rendle et al. 2009) and Optimizing GAUC (OptGAUC) (Song and Meyer 2015).

Parameter Setting. For all the above benchmark methods, we set the parameters recommended in the literature. For instance, we adopt λ=20 and r=50 in OptGAUC, while we set λ=1 and r=10 in LRM. As for the feature-based model CN, we use the difference between the number of positive and negative common neighbors as the metric to generate the ranking list.

In our FILE framework, there are three hyper-parameters: λ_1, λ_2 and r. Being consistent with the literature, we set λ_1=λ_2 and search over $\{0.01, 0.05, 0.1, 0.5\}$. We also search the number of latent features r over $\{5, 10, 15, 20\}$. We conduct 5 fold cross-validation on the training set and adopt the parameters which gain the best performance. We also check the parameter sensitivity of our approach with regard to λ_1, λ_2 and r, and the results on Slashdot@50 and Epinions@50 are presented in Figure 4. Across all parameters combinations, in terms of GAUC, FILE varies in a range of $[0.823, 0.856]$ in Slashdot@50 and $[0.779, 0.823]$ in Epinions@50. We can see that the performance fluctuation over different parameter settings is relatively small. We get similar results in other datasets as shown in Figure 5. The maximum fluctuation is 0.051 and occurs in Wikipedia@10. We can thus conclude that FILE shows good flexibility because of its insensitivity to the model parameters.

Comparative Experiments

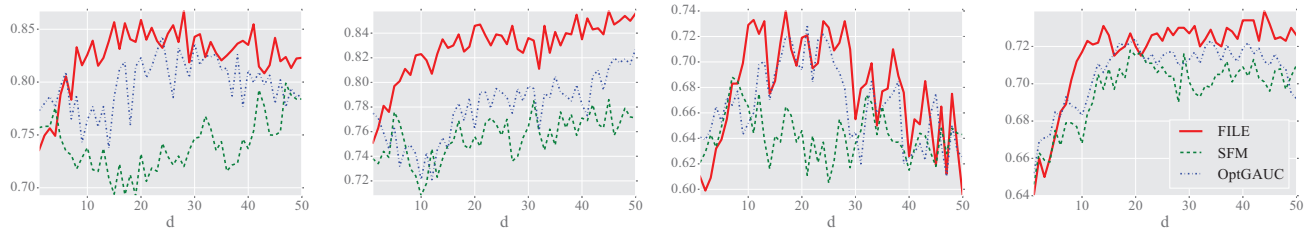
Overall Performance. Table 3 shows the comparisons among different models regarding to the ranking metric GAUC. As demonstrated, our model outperforms other benchmarks on most of the datasets. CN performs the worst in all scenarios because it does not differentiate the signs of neighbors and links, which indicating that traditional link prediction methods cannot be directly applied for link prediction in signed networks. The latent models, LRM, BPRMF and OptGAUC, perform better than CN, which shows the effectiveness of the latent features. In addition, OptGAUC outperforms LRM and BPRMF, indicating that no-relation information used in OptGAUC helps improve the performance of link prediction. This result is consistent with the result in (Song and Meyer 2015). Besides, SFM performs better than CN, LRM and BPRMF, suggesting that the explicit social features in SFM work well in signed network scenarios. In Wikipedia@50, FILE performs worse than OptGAUC and SFM, but the high variation (−8.07%) is caused by only a few mispredictions as Wikipedia@50 is a very small dataset. Besides, FILE improves its performance as more data is considered, i.e., in Wikipedia@10 and



(a) E@10 (b) E@25 (c) E@50 (d) S@10 (e) S@25 (f) S@50

(g) E@10 (h) E@25 (i) E@50 (j) S@10 (k) S@25 (l) S@50

Figure 6: (a)-(f) represent PPrec@top k; (g)-(l) refer to NPrec@ top k.



(a) Epinions (b) Slashdot (c) Wikipedia (d) Bitcoin

Figure 7: The impact of degree d.

Wikipedia@25. As suggested by SFM, the performance of FILE can be improved in the dataset like Wikipedia@50 by incorporating more explicit features.

Overall, FILE achieves the best performance when compared with other approaches across all the datasets, and the improvement is 3.9% on average. We conduct t-test for the performance difference over these approaches, and the result shows that the improvement of our framework is statistically significant (p-value < 0.01).

Top-k Ranking Performance. We investigate the ranking performance on top k. Both precisions of positive (i.e., PPrec) and negative (i.e., NPrec) at top k (k={10, 25, 50, 100, 200, 500, 1000}) are examined. Due to space limitations, we show the experimental results in six datasets for each metric in Figure 6, and the results are consistent in the other six datasets. For clarity, we only show the performance of OptGAUC and SFM, which perform better than the other three competing approaches. We can see that in terms of PPrec and NPrec, FILE consistently achieves the best results in almost all scenarios, demonstrating the usefulness of our approach since top-k list is very piratical and effective in real-world applications. It is worth noting that, FILE exhibits greater improvement over other approaches in term of NPrec, indicating its immense potential in security-related applications. The overall results again verify the effectiveness of incorporating latent and explicit features for link prediction in signed networks.

Impact of degree d. To demonstrate the robustness of our approach, we check its performance in terms of GAUC as the change of d in the range $[1, 50]$. As shown in Figure 7, when d is small ($d < 5$), FILE performs similarly to or slightly worse than others. As d increases, our approach is consistently better than others. One reason is that the data of larger d preserves more valuable information to learn latent features via matrix factorization. Besides, as we have mentioned, in reality d is usually much bigger than 5, we thus are more convinced of the robustness and superior of our approach in real-world scenarios. Another observation is that the performance falls as the degree increases in the Wikipedia dataset. It's because the Wikipedia dataset is very small. When the degree is larger than 20, the filtered dataset gets smaller. This is why the performance of all approaches becomes worse.

Conclusions

Link prediction in signed networks is challenging because of the imbalance of the three kinds of social status, which are positive, negative and no-relation. Besides, previous methods cannot well predict no-relation status due to the difficulty in distinguishing the no-relation of the stranger and frenemy types from the linked types. Therefore, in this paper, inspiring by the psychosocial theories, we propose the FILE framework which considers both social linkage criteria of individual users and the external social influence from the

neighborhood of every user pair. We also particularly design an optimization approach for this problem using the matrix factorization technique with a ranking-oriented loss function. Extensive evaluations in four datasets show that our model outperforms state-of-the-art approaches, demonstrating that our framework has effectively incorporated latent and explicit features for link prediction in signed networks. Besides, experimental results also verify that FILE is robust and relatively insensitive to the choice of model parameters.

In future, we will explore more explicit features for the FILE framework to enhance its prediction performance, and further test the effectiveness of it using field experiments.

Acknowledgments

This work is supported by the MOE AcRF Tier 1 funding (M4011261.020) and the Telenor-NTU Joint R&D funding awarded to Dr. Jie Zhang, and by the National Natural Science Foundations of China NSFC-71601104 and the Basic Academic Discipline Program for Shanghai University of Finance and Economics awarded to Dr. Hui Fang.

References

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

Al Hasan, M.; Chaoji, V.; Salem, S.; and Zaki, M. 2006. Link prediction using supervised learning. In *Sixth SIAM International Conference on Data Mining: Workshop on Link Analysis, Counter-terrorism and Security*.

Bukowski, W. M., and Hoza, B. 1989. Popularity and friendship: Issues in theory, measurement, and outcome.

Chiang, K.-Y.; Natarajan, N.; Tewari, A.; and Dhillon, I. S. 2011. Exploiting longer cycles for link prediction in signed networks. In *Proceedings of the 20th ACM International Conference on Information and knowledge management*, 1157–1162. ACM.

Duck, S. W., and Craig, G. 1978. Personality similarity and the development of friendship: A longitudinal study. *British Journal of Clinical Psychology* 17(3):237–242.

Express.co.uk. 2017. The average person has this many friends. https://goo.gl/bN47rq.

Hsieh, C.-J.; Chiang, K.-Y.; and Dhillon, I. S. 2012. Low rank modeling of signed networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 507–515. ACM.

Kumar, S.; Spezzano, F.; Subrahmanian, V.; and Faloutsos, C. 2016. Edge weight prediction in weighted signed networks. In *Proceedings of the IEEE 16th International Conference on Data Mining*, 221–230. IEEE.

Leskovec, J.; Huttenlocher, D.; and Kleinberg, J. 2010a. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*, 641–650. ACM.

Leskovec, J.; Huttenlocher, D.; and Kleinberg, J. 2010b. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1361–1370. ACM.

Li, X.; Fang, H.; and Zhang, J. 2017a. A feature-based approach for the redefined link prediction problem in signed networks. In *Proceedings of the 13th International Conference on Advanced Data Mining and Applications*.

Li, X.; Fang, H.; and Zhang, J. 2017b. Rethinking the link prediction problem in signed social networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4955–4956.

Liben-Nowell, D., and Kleinberg, J. 2007. The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology* 58(7):1019–1031.

Lichtenwalter, R. N., and Chawla, N. V. 2012. Vertex collocation profiles: subgraph counting for link analysis and prediction. In *Proceedings of the 21st International Conference on World Wide Web*, 1019–1028. ACM.

Lichtenwalter, R. N.; Lussier, J. T.; and Chawla, N. V. 2010. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 243–252. ACM.

Mazie, S. 2016. Do you have too many facebook friends? https://goo.gl/zkLTfe.

Menon, A. K., and Elkan, C. 2011. Link prediction via matrix factorization. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 437–452. Springer.

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

Nguyen, T.; Phung, D. Q.; Adams, B.; and Venkatesh, S. 2011. Towards discovery of influence and personality traits through social link prediction. In *Fifth International AAAI Conference on Weblogs and Social Media*, 566–569.

Papaoikonomou, A.; Kardara, M.; Tserpes, K.; and Varvarigou, T. A. 2014. Predicting edge signs in social networks using frequent subgraph discovery. *IEEE Internet Computing* 18(5):36–43.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 452–461. AUAI Press.

Song, D., and Meyer, D. A. 2015. Recommending positive links in signed social networks by optimizing a generalized auc. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 290–296.

Song, D.; Meyer, D. A.; and Tao, D. 2015. Efficient latent link recommendation in signed networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1105–1114. ACM.

Tang, J.; Chang, Y.; Aggarwal, C.; and Liu, H. 2016. A survey of signed network mining in social media. *ACM Computing Surveys (CSUR)* 49(3):42.

Ye, J.; Cheng, H.; Zhu, Z.; and Chen, M. 2013. Predicting positive and negative links in signed social networks by transfer learning. In *Proceedings of the 22nd International Conference on World Wide Web*, 1477–1488. ACM.

Zhou, T.; Lü, L.; and Zhang, Y.-C. 2009. Predicting missing links via local information. *The European Physical Journal B-Condensed Matter and Complex Systems* 71(4):623–630.