A Feature-Based Approach for the Redefined Link Prediction Problem in Signed Networks

Xiaoming Li[1], Hui Fang[2](✉), and Jie Zhang[1]

[1] School of Computer Science and Engineering, Nanyang Technological University,
Singapore, Singapore
{lixiaoming,zhangj}@ntu.edu.sg

[2] Shanghai University of Finance and Economics, Shanghai, China
fang.hui@mail.shufe.edu.cn

Abstract. Link prediction is an important research issue in social networks, which can be applied in many areas, such as trust-aware business applications and viral marketing campaigns. With the rise of signed networks, the link prediction problem becomes more complex and challenging as it introduces negative relations among users. Instead of predicting future relation for a pair of users, however, the current research focuses on distinguishing whether a certain link is positive or negative, on the premise of the link existence. The situation that two users do not have relation (i.e., no-relation) is also not considered, which actually is the most common case in reality. In this paper, we redefine the link prediction problem in signed social networks by also considering "no-relation" as a future status of a node pair. To understand the underlying mechanism of link formation in signed networks, we propose a feature framework on the basis of a thorough exploration of potential features for the newly identified problem. We find that features derived from social theories can well distinguish these three social statuses. Grounded on the feature framework, we adopt a multiclass classification model to leverage all the features, and experiments show that our method outperforms the state-of-the-art methods.

Keywords: Signed social network · Link prediction · No-relation

1 Introduction

Signed network, literally denotes the network which contains both positive and negative links among nodes. Under this network structure, the relationship between online users can be 'friend' or 'foe', and 'trust' or 'distrust', etc. The rise of this new type of user relationship networks has broad implications for real businesses nowadays, and more and more online systems such as Slashdot (friend or foe, www.slashdot.org), Epinions (trust or distrust, www.epinions.com) and even Facebook have adopted signed network structure and features. For example, Facebook introduced a handful of new reaction buttons, besides 'like' option, to show different attitudes such as 'angry' and 'sad' to other users. In other words,

G. Cong et al. (Eds.): ADMA 2017, LNAI 10604, pp. 165–179, 2017.
https://doi.org/10.1007/978-3-319-69179-4_12

relationship between online users are not only limited to positive (e.g. friend and trust) anymore, but try to add more alternatives to be consistent with the human relationship in real life.

The increasing interest in signed social networks has brought great impact on many traditional research topics, one of which is link prediction. Link prediction, which aims to infer the formation of a possible link in the near future, is well studied in the last few years as its significant contributions to improve and enhance online experiences [11,14], in the form of further facilitating applications such as recommenders for products or friends [22] and social networks [23]. Link prediction in unsigned networks aims to predict the future connection status between two nodes, or a dyad, either linked or not. On the contrary, connection status of two nodes in signed networks could be positive, negative and no-relation, which increases the difficulty of link prediction.

On the other hand, link sign prediction in signed networks focuses on predicting signs of existing links [4,7,9,18], which is a binary classification problem. In other words, this kind of research basically ignores the no-relation status. The rationale behind this assumption might be two parts: (1) if no-relation is considered as a link status, positive or negative ones will be highly imbalanced and sparse contrasted with no-relation, and thus most of the machine learning methods incline to predict the dyad status to be no-relation for the sake of maximum accuracy; (2) for some traditional applications (e.g. spam email detection), the assumption of link existence is inherently satisfied.

However, most link prediction applications in signed networks cannot be simply treated as the sign prediction problem as aforementioned. For example, in voting prediction, a user might vote a candidate entity as positive or negative, but in most cases, the user will choose not to vote the entity. Therefore, the existing methods cannot be directly adopted to address these kinds of applications. To conclude, the previous approaches mainly suffer from two issues: (1) they ignore the no-relation status, which accounts for the majority of the real relationship in signed social networks; (2) instead of predicting the future relationship status of any two nodes, these approaches actually consider a static network as they assume the existence of links with uncertain signs.

In this paper, we design a link prediction approach for signed social networks, to predict future link status, which could be positive, negative or no-relation. We take no-relation and future status into consideration, which are the two major differences compared with the previous studies in the literature. To address the problem, we first thoroughly explore features which may potentially affect future link status of any two nodes, especially to investigate the related features which can well distinguish no-relation from the other two statuses. On the basis of the thorough feature exploration, we propose a feature framework, where features of different categories try to distinguish the three link statuses, for link prediction in signed networks, and design a simple but effective feature selection mechanism to show how to apply the feature framework in real applications. With the feature framework, we establish a feature based link prediction model in signed networks. Experiments verify the effectiveness of the proposed feature framework, and demonstrate that our model outperforms the state-of-the-art approaches for both the measurements of Positive AUC and Generalized AUC [18].

To summarize, the main contributions of our work are two-folds:

1. We redefine the link prediction problem by taking an initial step to consider 'no-relation' as a future dyad status for link prediction in signed networks. Besides, we focus on predicting the future relationship status of any two nodes, rather than distinguishing the sign of a certain link in a static network, which is the common setting of the current approaches [4,7,9,18].
2. We propose a structured feature framework for the redefined problem on the basis of a thorough feature analysis to reveal the underlying mechanism regarding link formation in signed networks. The feature framework, grounded on both well-known theories and sound observations, can serve as a guidance for research on the new problem.

2 Related Work

We summarize the literature into two parts: (1) link prediction in unsigned networks; and (2) link prediction in signed networks.

Link prediction in unsigned networks has been well studied during the past decade. Existing methods can be divided into two classes: unsupervised and supervised ones. Unsupervised methods consist of neighbor-based metrics and path-based metrics [11]. They calculate a "link formation score" for each pair of nodes to indicate their possibility to be linked or not in the near future. Popular neighbor-based metrics include: the number of common neighbors [1], Adamic/Adar Index [1], Jaccard Coefficient [11], Preferential attachment [16] and Resource Allocation Index [24]. These ranking metrics are derived from neighborhood structure. Meanwhile, the features related to the path between two nodes in a network structure are also used to compute the similarities of node pairs, like Katz [8], Vertex Collocation Profile [12] and ProfFlow [13]. Popular supervised methods include: feature-based classification models [2], and latent feature models [15]. However, link prediction in unsigned networks considers only two possible future connection statuses of two nodes, i.e., linked or not-linked, while in signed networks, three connection statuses are possible: positive, negative, and no-relation. Therefore, all the features and metrics need to be re-investigated in the signed network scenario, because neighbors and paths can be negative in signed networks.

Existing attempts for signed link prediction mainly focus on how to distinguish positive and negative links, and topology feature-based approaches are dominant in the literature. For example, based on balance theory and status theory, Leskovec et al. [9] identify triangle-based features of each two users and their common neighbours to predict the sign (i.e., positive or negative) between each two users. Besides, k-cycle-based features are proposed in [4] 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. [17] leverage the pattern of frequent subgraph among node pairs, to predict link status. Another type of popular methods is the low-rank models. For example, Hsieh et al. [7] verify that signed networks naturally present a low-rank structure, and a matrix factorization model is proposed to infer link signs. However, all the aforementioned methods assume the existence of links with uncertain signs. Song and Meyer [18] also adopt a low-rank model to infer link signs, which learns the latent features by minimizing the generalized AUC loss. Although it takes the no-relation into consideration, the major purpose of the work is to distinguish positive and negative link status, and the no-relation information is only used in the training period of the model.

We can see that the no-relation status is ignored in the existing studies, while no-relations actually account for the majority of the real relationship in signed social networks. Besides, we need to re-investigate the features adopted in the literature for link prediction. We then can design more specific approaches for the newly redefined problem (i.e., link prediction in signed network) on the basis of a better feature design and problem analysis.

3 Redefined Link Prediction Problem

Here we first formalize our redefined link prediction problem in signed social networks. Specifically, let $G = (V, E^P, E^N, X)$ denote a signed social network, where V is the node set; E^P is the set of positive links and E^N is the set of negative links; X refers to the set of no-relation. $G_t = (V, E_t^P, E_t^N, X_t)$ denotes the snapshot of the network at time t. Our research question is: *given a series of network snapshots G_0, G_1, ..., G_t, and any node pair (i, j) (i.e. dyad) where $x_{ij} \in X_t$, to predict the connection status of x_{ij} at time $t + 1$, which can belong to E_{t+1}^P, E_{t+1}^N or X_{t+1}.*

To be more specific, in this paper, we aim to solve three questions as below:

1. What has been changed with the introduction of no-relation?
2. What is the link formation mechanism behind signed network evolution? or which specific features influence link formation in signed networks?
3. How to evaluate link prediction performance involving no-relation?

To address these questions, we propose a link prediction approach in signed networks, including a feature framework of six categories, a feature-based link prediction model, and a feature selection mechanism.[1] We also introduce two techniques to address the data imbalance issue for link prediction in signed networks.

[1] The preliminary version [10] of our work has been published at AAAI 2017 as a student abstract.

4 Feature Framework

4.1 Data Description

We first introduce the datasets used in this work. We obtain two publicly available datasets[2] with the signed structure, i.e., Epinions and Slashdot. In both social networks, users can establish trust and distrust relationship, i.e., positive or negative links with other users. Table 1 provides descriptive statistics for these datasets.

Table 1. Dataset statistics

	Epinions	Slashdot
Users	$131,828$	$82,140$
Users with degree ≥ 10	$17,664$	$17,794$
Users with degree ≥ 25	$9,134$	$8,325$
User pairs with positive links	$717,667$	$425,072$
User pairs with negative links	$123,705$	$124,130$
User pairs with no-relations	$1.7 * 10^{10}$	$6.7 * 10^{9}$

From Table 1, we can quickly summarize two general data patterns that occur in signed networks: (1) sparsity: a signed network is quite sparse, as there are no more than 10.2% users with degree ≥ 25 in both real social networks; and (2) imbalance: the number of linked pairs is smaller than pairs of no-relation by four orders of magnitude. Meanwhile, the number of positive and negative links are also imbalanced.

4.2 Feature Design Principles

The design of a feature set is always the keystone of a feature-based prediction method. Link prediction in unsigned networks adopts features, such as the number of common friends, to distinguish linked and not-linked status corresponding to those with the value of 1 and 0 in unsigned networks. On the contrary, previous link prediction in signed networks designs features to discriminate positive links and negative links, with the value of 1 and -1 respectively. In our newly identified link prediction problem in signed networks, as three link statuses (with the value of 1, 0 and -1) are involved, the feature set should be re-considered.

An ideal feature is expected to well distinguish the three link statuses, however, as indicated before, there is no previous feature study considering the three statuses together. To fill this gap, we propose a feature framework for the new research problem, aiming to serve as a guidance and elicit more related research. We not only adopt existing features in previous studies [9,20] on both unsigned and signed network scenarios, but also derive new features based on our analysis

[2] https://snap.stanford.edu.

and observations. We then combine and summarize these features into six major categories, and then explore the influence of each category on link formation in signed networks. We also indicate how our features have addressed our problem uniquely. All features are discussed in the following section.

4.3 Feature Definition

Balance Theory [3] can be simply explained as "my friend's friend is my friend", or "my enemy's enemy is my friend". In other words, two users will more likely become friends if they have many common friends. Thus, we define **pp** and **pp_ratio** to represent the number and the fraction of the common 'positive' neighbors (friends) between two users, and **nn** and **nn_ratio** to represent the number and the fraction of the 'negative' neighbors (enemies). Besides, given two users, we also check the number of their neighbors which are one's friends but the other's enemies, denoted by **pn**. Based on balance theory, a large **pn** represents a high chance for a negative link establishment. Then we define a feature **bal_diff** to check the contradiction within the balance theory. When user i and j have largely the same number of 'friends' and 'enemies', a positive or negative sign will eventually make the network unbalanced, on the basis of social balance theory. In this research, balance theory is extended such that the no-relation status can make the graph more balanced.

Status Theory [5,9] refers to that, a positive link $i \rightarrow j$ indicates the node status of j is higher than i. Therefore, given a common neighbor w, if link $i \rightarrow w$ and $w \rightarrow j$ are both positive, link $i \rightarrow j$ is more likely to be positive since the status of j is higher than i. Thus, given two users i and j, we define **sta_diff_p** (**sta_diff_n**) as the number of their neighbors which indicate j's status is higher (lower) than i. Then **sta_diff** is used to represent the status difference between these two users, while **sta_diff_ratio** takes into account the fraction of the status difference. Status theory is extended in this research, as two users tend to have no-relation if they have nearly equal status.

Reciprocity [6] is the tendency that two nodes with bidirectional links between each other always have the same sign. In Epinions dataset, 83.5% of user pairs with bidirectional links have the same sign. Therefore, we can infer the status of the link $i \rightarrow j$ by the sign of the backward link $j \rightarrow i$, named as **reciprocity**. This feature will be useful if there exist many bidirectional links in the network.

Rich-get-richer [21] indicate two active or popular users will more likely get linked. Thus, we derive 10 features to capture this phenomenon. Given two user i and j, we define **out_p** and **out_p_ratio** to represent the number and the fraction of positive links coming from i. Similarly, **out_n** takes into account the negative links. Meanwhile, two features **in_p** and **in_p_ratio** are the number and the fraction of the positive links pointed to j. Besides, if i's out_p_ratio and j's in_p_ratio are both high, there will be more likely a positive link between i and j. However, if i's out_p_ratio (or out_n_ratio) is large and the j's in_n_ratio (or in_p_ratio) is large, which indicates that i is active and tends to trust others, but j is not trustworthy and distrusted by others, there will be no-relation between i and j. Therefore, we adopt 4 features **prprs**, **prnrs**, **nrprs**, **prnrs** to capture those observations and check whether those features can indicate no-relation status.

Clustering [11] adopts the similar insight with link prediction in unsigned networks. It measures per-dyad side features like the number of common neighbors. The underlying assumption is that two users likely get connected if they have many common neighbors. In this work, we use 5 features **CN** [11], **Katz** [8], **JC** [11], **PA** [16], **Status Similarity** [19]. In signed network, a smaller feature in this category indicates a higher chance to have no-relation.

Frequent Subgraph [4,9] considers triads constructed by users i, j and their common neighbors. Each link between a user and its neighbor may have two directions, i.e., forward and backward, meanwhile it can be positive or negative. Therefore, based on the combination of the directions and signs, there will be 16 types of triads. We use p and n to represent the positive and negative signs, and f and b denote the link direction, so these triads represent as $ppff$, $pnfb$, etc., as shown in Fig. 1.

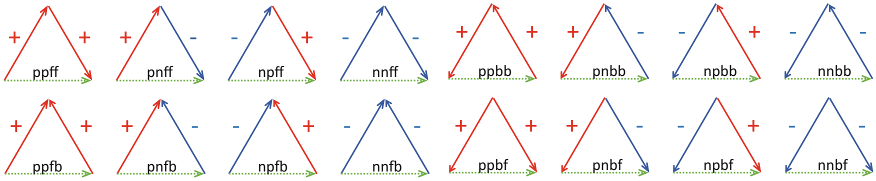


Fig. 1. The sixteen triads are fundamental and crucial units for network topology analysis [4,9].

In summary, balance theory, status theory and reciprocity mainly capture the signed network characteristics; rich-get-richer considers per-node side features meanwhile clustering captures per-dyad side features; and frequent subgraph captures relatively larger scale topological features.

We adopt the following notation: let $1, 2, \cdots, N$ be N users; Let S_{ij} be the link sign from user i toward user j; Let O_i, I_i be user i's outgoing and incoming link sets, respectively. Specifically, O_i^+, I_i^+ represent the positive link sets, and O_i^-, I_i^- the negative link sets; Let C_{ij} be the set of the common neighbors between users i and j; ppff, pnff, ... are basic triad units. Table 2 summarizes the full list of the features we derived from social theories.

Table 2. Features derived from social theories

Feature	Notation
Balance theory	
pp	ppff + ppfb + ppbf + ppbb
nn	nnff + nnfb + nnbf + nnbb
pn	$\lvert C\rvert-$ pp $-$ nn
pp_ratio	pp/$\lvert C\rvert$
nn_ratio	nn/$\lvert C\rvert$
bal_diff	pp + nn $-$ pn
Status theory	
sta_diff	sta_diff_p $-$ sta_diff_n
sta_diff_p	ppff + nnbb + pnfb + npbf
sta_diff_n	nnff + ppbb + npfb + pnfb
sta_diff_ratio	sta_diff_p/(sta_diff_p + sta_diff_n)
Reciprocity	
Reciprocity	S_{ji}
Rich-get-richer	
out_p	$\lvert O_i^+\rvert$
out_n	$\lvert O_i^-\rvert$
in_p	$\lvert I_j^+\rvert$
in_n	$\lvert I_j^-\rvert$
out_p_ratio	$\lvert O_i^+\rvert/\lvert O_i\rvert$
in_p_ratio	$\lvert I_j^+\rvert/\lvert I_j\rvert$
prprs	$(\lvert O_i^+\rvert/\lvert O_i\rvert) * (\lvert I_j^+\rvert/\lvert I_j\rvert)$
prnrs	$(\lvert O_i^+\rvert/\lvert O_i\rvert) * (\lvert I_j^-\rvert/\lvert I_j\rvert)$
nrnrs	$(\lvert O_i^-\rvert/\lvert O_i\rvert) * (\lvert I_j^-\rvert/\lvert I_j\rvert)$
nrprs	$(\lvert O_i^-\rvert/\lvert O_i\rvert) * (\lvert I_j^+\rvert/\lvert I_j\rvert)$
Clustering	
cn	$\lvert C\rvert$
Katz	ppff + pnff + npff + nnff
Jaccard coefficient	$\lvert C_{ij}\rvert/(\lvert O_i\rvert + \lvert O_j\rvert + \lvert I_i\rvert + \lvert I_j\rvert)$
Preferential attachment	$(\lvert O_i\rvert + \lvert I_i\rvert) * (\lvert O_j\rvert + \lvert I_j\rvert)$
Status similarity	$1/(\delta(i) + \delta(j) - 1)$, $\delta(i) = \lvert I_i^+\rvert + \lvert O_i^-\rvert - \lvert O_i^+\rvert - \lvert I_i^-\rvert$

5 Signed Link Prediction

The proposed feature-based model can be stated as:

$$\min_{\alpha,\beta} \sum l(S_{ij}, L(\alpha f(u_i, u_j) + \beta u_{ij})) + \frac{\lambda_1}{2}\|\alpha\|_2^2 + \frac{\lambda_2}{2}\|\beta\|_2^2 \qquad (1)$$

where S_{ij} is the ground truth of link status; $\mathrm{L}(\cdot)$ is the link prediction function; $l(\cdot,\cdot)$ is the loss function; u_{ij}, u_i and u_j are corresponding features; $\frac{\lambda_1}{2}\|\alpha\|_2^2$ and $\frac{\lambda_2}{2}\|\beta\|_2^2$ are regularizers.

Link prediction function $\mathrm{L}(\cdot)$ is a function with a value of 1, -1 or 0, which represents positive, negative or no-relation respectively. Under this setting, a multiclass classification algorithm should be adopted, such as SVM and decision trees. In this paper, we adopt the multinomial logistic regression model.

Loss function $l(\cdot,\cdot)$ is user-specified and application-depended. For example, in recommendation systems, loss for an incorrectly predicted -1 or 0 can be relatively low, while the loss for a mistakenly identified 1 should be set high, as the prediction performance on 1 is of the most importance.

Features mainly consist of two parts: per-dyad side, u_{ij} is the feature set of dyad (i,j), such as **sta_diff**, **pp**, **pp_ratio**; per-node side, $f(u_i, u_j)$ is the function to leverage the node-side features of u_i and u_j, like **prprs**, which is the multiplication of i's **out_p_ratio** and j's **in_p_ratio**.

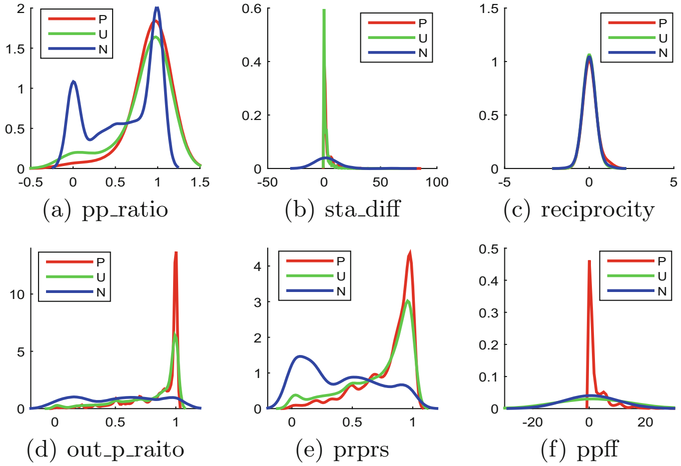
5.1 Feature Selection Mechanism

Before using these potential predictive features, we need to investigate whether these features can distinguish different classes, or have different influences on each class. As aforementioned, an ideal feature is expected to well distinguish the three link statuses. For each specific application, to effectively adopt the feature framework, we should firstly investigate whether each theoretically sound feature is suitable for the real application.

To do this, we statistically check the mean of each feature for each class (i.e. positive, negative or no-relation), taking Epinions dataset as an example. We conduct One-Way ANOVA test on $M_1(f_i)$, $M_0(f_i)$ and $M_{-1}(f_i)$, where f_i denotes a feature and $M(\cdot)$ denotes its average value. The corresponding null hypothesis is: $H_0 : M_1(f_i) = M_0(f_i) = M_{-1}(f_i)$. If a feature is rejected at the significance level of $\alpha = 0.01$ with p-value < 0.001, the feature is dropped in this application. We choose a smaller significance level and p-value here in order to strongly support the alternate hypothesis, i.e., try to select the features which can better distinguish those three classes. Figure 2 shows the kernel smoothed density distribution of some selected features. As shown in Fig. 2, we can easily understand why these features work. For example, in the figure, "prprs", the multiplication of node i's outgoing positive link ratio and node j's incoming positive link ratio, shows totally different distributions for different link status.

5.2 Handling Imbalance Issue

The imbalance issue is one of the most serious problems for link prediction in signed networks, where the number of no-relation pairs $\gg$ the number of positive links $\gg$ the number of negative links, as shown in Table 1. Therefore, if we conduct experiments based on the full dataset, positive and negative links will be almost ignored as there are overwhelmingly no-relation pairs. Meanwhile,



(a) pp_ratio (b) sta_diff (c) reciprocity (d) out_p_raito (e) prprs (f) ppff

Fig. 2. Kernel smoothed density distribution of selected features. Details of features are given in Table 2. P, U and N in legends represent positive, no-relation and negative link respectively. [Figures are best viewed in color]

the accuracy performance can reach to almost 100% since the learning model can predict all pairs as no-relation.

Thus, the first technique is under sampling, where we randomly draw a set of links including an equal amount of samples for each link status. Specifically, since the negative link is the smallest in quantity, for every negative link, we randomly draw a positive link and a no-relation pair.

Another technique is to use the measurements of ranking rather than the accuracy metric, to evaluate the prediction performance of different methods. As we have three statuses in signed networks, we aim to rank user pairs based on the predicted link scores, and make positive links be ranked higher at the top, and negative links lower in the bottom of the list. That is, the ranking order comes as positive, no-relation, negative ones in the list. In this work, we adopt GAUC (Generalized AUC) [18], an extension of AUC, as a metric which can measure the ranking performance for three statuses. A score of 1.0 indicates a perfect classifier while 0.5 represents a random classifier. This metric is insensitive to the imbalanced data.

6 Experiments

In this section, we conduct experiments using Epinions and Slashdot datasets to demonstrate the effectiveness of our feature framework, and the superiority of our link prediction approach compared to the state-of-the-art methods.

6.1 Experimental Setting

First, we aim to design an experimental environment which can well represent the link prediction scenario in reality. One realistic scenario is that, given a certain number of user pairs which currently are not linked, we predict which user pair will form a positive link or a negative link, or still have no-relation in the future. Since Epinions dataset contains a timestamp for each generated link, we can use it to test the performance of our method on future link prediction. We divide the dataset into three parts by timestamps: $T1$, $T2$ and $T3$, which represent the past, current and future respectively. Because there are $578,996$ links marked with the timestamp 1/10/2001, we treat this timestamp as "past" and use those links to derive training features. And we split the rest into two parts: training set consists of the links (or user pairs) formed during $T2$ (till 4/30/2002); and the testing set includes the links formed in the period of $T3$ (till 8/12/2003) but the features are measured in both periods of $T1$ and $T2$. Although there is an overlap between the feature sets of the training and testing data, this experiment setting is exactly consistent with the training and prediction process in the real-world scenarios. Based on the undersampling method discussed in Sect. 5.2, we sample a positive link and a no-relation dyad for every negative link, to ensure the training and testing data are balanced. Specifically, the number of samples in $T2$ and $T3$ are 18489 and 15741 respectively.

Since there is no timestamp information in the Slashdot dataset, we adopt the traditional training/testing setting [7,9], i.e., by randomly drawing a sample of user dyads including positive, negative and no-relation ones with the equal amount as training and testing set respectively. We adopt 10-fold cross validation for this dataset. Besides, to measure the effectiveness and robustness of our approach, we test our approach under different settings. We filter user pairs by different number of common neighbors, i.e., minimum as 1, 10 and 25. In the following experiments, if not stated otherwise, we show the result on user pairs with at least 1 neighbor since this is a more general setting.

Evaluation Metrics. As discussed in Sect. 5.2, we will use measurements of ranking rather than of accuracy. We adopt the generalized AUC (i.e., GAUC) metric [18], which is defined as:

$$\begin{aligned} GAUC = & \frac{1}{|P|+|N|}\Big(\frac{1}{|U|+|N|}\sum_{a_i \in P}\sum_{a_s \in U \cup N} I(L(a_i) > L(a_s)) \\ & + \frac{1}{|P|+|N|}\Big(\frac{1}{|U|+|P|}\sum_{a_j \in N}\sum_{a_t \in U \cup P} I(L(a_j) < L(a_t)) \end{aligned}$$

where $|P|$, $|N|$, $|U|$ represent the number of positive links, negative links, and no-relations, respectively; a represents a link; and L$(\cdot)$ is the link score function. GAUC is an extension of AUC, and provides a ranking metric considering the three link statuses.

The other metric is PAUC (positive AUC), which measures the classification performance over positive links and non-positive links. We do not show NAUC (negative AUC) results here since NAUC results can be derived from GAUC and AUC, where GAUC can be treated as the weighted sum of PAUC and NAUC.

Benchmark Approaches. We compare our approach with the following methods:

- Common Neighbors (CN) [11], ranks user pairs by the number of their common neighbors, including both the positive and negative common neighbors.
- Katz [8], ranks user pairs by the number of directional routes. CN and Katz are always used as the baselines for link prediction.
- Triad and Degree feature-based method(All23) [9], adopts total 23 features based on topolgy triads and user degree, and use regression as the learning model.
- Matrix Factorization (MF) [7], learns latent features from the social matrix with non-zero elements, and rank user pairs by the multiplication of latent features. This is a point-wise approach for sign prediction.
- Optimizing GAUC (OptGAUC) [18], optimizes the GAUC metric through matrix factorization with the ranking order of the positive, no-relation, and negative links. It is a pair-wise approach for sign prediction.

6.2 Prediction Performance

As shown in Table 3, we can see that our approach outperforms others on both GAUC and PAUC metrics, under different dataset settings.

Table 3. Performance comparison

	Epinions						Slashdot					
Method	cn $\geq$ 1		cn $\geq$ 10		cn $\geq$ 25		cn $\geq$ 1		cn $\geq$ 10		cn $\geq$ 25	
	GAUC	PAUC	GAUC	PAUC	GAUC	PAUC	GAUC	PAUC	GAUC	PAUC	GAUC	PAUC
CN	0.576	0.587	0.566	0.57	0.545	0.556	0.625	0.649	0.643	0.697	0.645	0.699
Katz	0.591	0.592	0.602	0.571	0.549	0.55	0.661	0.665	0.697	0.752	0.712	0.758
MF	0.654	0.645	0.657	0.651	0.662	0.658	0.552	0.545	0.565	0.558	0.561	0.559
OptGAUC	0.715	0.72	0.709	0.702	0.719	0.712	0.603	0.599	0.613	0.601	0.619	0.605
Triads + Degree	0.742	0.736	0.825	0.777	0.838	0.798	0.878	0.853	0.887	0.862	0.892	0.865
Ours	**0.827**	**0.799**	**0.834**	**0.791**	**0.840**	**0.807**	**0.923**	**0.904**	**0.924**	**0.897**	**0.926**	**0.898**

CN and Katz do not perform well because they do not differentiate the signs of neighbors and links. Thus we conclude that traditional link prediction methods cannot directly be applied for link prediction in signed networks. OptGAUC outperforms MF, indicating that no-relation information used in OptGAUC helps improve its link prediction. As the PAUC measurement can be treated as an extended version of AUC in traditional unsigned networks, we can thus conclude that negative links as new information for link prediction can improve the performance of predicting positive links in signed networks.

6.3 Feature Framework Analysis

In order to check the effectiveness of each feature category and the robustness of the feature framework, we first check the prediction performance of each feature category. As shown in Table 4, the learning model with any feature category outperforms random guessing (GAUC = 0.5). Specifically, the best performance in terms of both GAUC and PAUC is given by balance theory and frequent subgraph for Epinions dataset, meanwhile, cluster and frequent subgraph outperform others in Slashdot dataset. The full model which adopts all features shows the best performance, demonstrating the effectiveness of feature combinations.

Table 4. The effectiveness of each feature category

Feature category	Epinions		Slashdot	
	GAUC	PAUC	GAUC	PAUC
Balance theory	0.733	0.734	0.808	0.794
Status theory	0.72	0.703	0.647	0.639
Reciprocity	0.538	0.549	0.622	0.645
Rich-get-richer	0.738	0.715	0.691	0.682
Cluster	0.617	0.638	0.820	0.807
Frequent subgraph	0.799	0.776	0.843	0.825
Full model	**0.827**	**0.799**	**0.924**	**0.904**

Furthermore, we evaluate the performance of our approach by removing the features of a certain category each time. The experimental results are shown in Table 5, where each row represents the prediction results in terms of GAUC and PAUC after dropping features of the corresponding category. We can see that the performance of each incomplete framework is worse than the complete one involving the features of all the six categories.

Table 5. The effectiveness of the framework by removing one feature category

Feature category	Epinions		Slashdot	
	GAUC	PAUC	GAUC	PAUC
Balance theory	0.801	0.774	0.921	0.902
Status theory	0.818	0.789	0.902	0.872
Reciprocity	0.822	0.791	0.905	0.874
Rich-get-richer	0.816	0.795	0.921	0.903
Cluster	0.823	0.798	0.875	0.867
Frequent subgraph	0.824	0.796	0.916	0.895
Full model	**0.827**	**0.799**	**0.924**	**0.904**

6.4 Link Prediction Model Comparison

We also examine the performance of different multiclass classifiers for the link prediction function in our feature-based approach. This test is conducted within the WEKA framework, and each model adopts the default parameter setting. Experimental results in Table 6 indicate that our approach with the multinomial logistic regression model achieves the best performance in term of GAUC.

Table 6. GAUC performance on different multiclass classification models

Classifier	Epinions	Slashdot
SVM	0.497	0.583
Decision tree	0.679	0.865
Adaboost	0.684	0.794
Naïve Bayes	0.796	0.866
Random forest	0.81	0.798
Multinomial logistic regression	**0.827**	**0.924**

7 Conclusions and Future Work

In this paper, we redefine the link prediction problem in signed networks, by considering no-relation as a future status of a user pair. For this problem, we further propose a feature framework grounded on thorough theoretical analysis, and design a feature selection mechanism and feature-based prediction model to apply the framework in real applications. We also indicate two techniques to handle the imbalance issue for link prediction in signed networks. Experiments in Epinions and Slashdot dataset show that our model outperforms existing methods in terms of GAUC and PAUC, and also demonstrate that each category of our feature framework and our choice of the multinomial logistic regression model are effective.

This work takes an initial step to consider 'no-relation' as a future status for link prediction in signed networks, and our proposed feature framework can serve as a leading guidance for research on the new problem. For future work, firstly, we will investigate more real-world datasets to further evaluate the significance of the new problem and the effectiveness of our approach. Secondly, we will explore more features and design an advanced model specifically for link prediction in signed networks.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Grant No. 71601104).

References

1. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Soc. Netw. **25**(3), 211–230 (2003)
2. Al Hasan, M., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: SDM (2006)
3. Antal, T., Krapivsky, P.L., Redner, S.: Social balance on networks: the dynamics of friendship and enmity. Phys. D Nonlinear Phenom. **224**(1), 130–136 (2006)
4. Chiang, K.Y., Natarajan, N., Tewari, A., Dhillon, I.S.: Exploiting longer cycles for link prediction in signed networks. In: CIKM, pp. 1157–1162. ACM (2011)
5. Davis, J.A., Leinhardt, S.: The structure of positive interpersonal relations in small groups (1967)
6. Falk, A., Fischbacher, U.: A theory of reciprocity. Games Econ. Behav. **54**(2), 293–315 (2006)
7. Hsieh, C.J., Chiang, K.Y., Dhillon, I.S.: Low rank modeling of signed networks. In: SIGKDD, pp. 507–515. ACM (2012)
8. Katz, L.: A new status index derived from sociometric analysis. Psychometrika **18**(1), 39–43 (1953)
9. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. In: WWW, pp. 641–650. ACM (2010)
10. Li, X., Fang, H., Zhang, J.: Rethinking the link prediction problem in signed social networks. In: AAAI (2017)
11. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. JAIST **58**(7), 1019–1031 (2007)
12. Lichtenwalter, R.N., Chawla, N.V.: Vertex collocation profiles: subgraph counting for link analysis and prediction. In: WWW, pp. 1019–1028. ACM (2012)
13. Lichtenwalter, R.N., Lussier, J.T., Chawla, N.V.: New perspectives and methods in link prediction. In: SIGKDD, pp. 243–252. ACM (2010)
14. Lü, L., Zhou, T.: Link prediction in complex networks: a survey. Phys. A Stat. Mech. Appl. **390**(6), 1150–1170 (2011)
15. Menon, A.K., Elkan, C.: Link Prediction via Matrix Factorization. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011. LNCS, vol. 6912, pp. 437–452. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23783-6_28
16. Newman, M.E.: Clustering and preferential attachment in growing networks. Phys. Rev. E **64**(2), 025102 (2001)
17. Papaoikonomou, A., Kardara, M., Tserpes, K., Varvarigou, T.A.: Predicting edge signs in social networks using frequent subgraph discovery. IEEE Internet Comput. **18**(5), 36–43 (2014)
18. Song, D., Meyer, D.A.: Recommending positive links in signed social networks by optimizing a generalized auc. In: AAAI, pp. 290–296 (2015)
19. Symeonidis, P., Tiakas, E.: Transitive node similarity: predicting and recommending links in signed social networks. WWW **17**(4), 743–776 (2014)
20. Tang, J., Chang, Y., Liu, H.: Mining social media with social theories: a survey. ACM SIGKDD Explor. Newsl. **15**(2), 20–29 (2014)
21. Tufekci, Z.: Who acquires friends through social media and why? "rich get richer" versus "seek and ye shall find". In: ICWSM (2010)
22. Zhang, J., Lv, Y., Yu, P.: Enterprise social link recommendation. In: CIKM, pp. 841–850. ACM (2015)
23. Zhao, T., Zhao, H.V., King, I.: Exploiting game theoretic analysis for link recommendation in social networks. In: CIKM, pp. 851–860. ACM (2015)
24. Zhou, T., Lü, L., Zhang, Y.C.: Predicting missing links via local information. EPLB-Condens. Matter Complex Syst. **71**(4), 623–630 (2009)