Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

# Rethinking the Link Prediction Problem in Signed Social Networks

**Xiaoming Li, Hui Fang,* Jie Zhang**
School of Computer Science and Engineering, Nanyang Technological University, Singapore
*Shanghai University of Finance and Economics, China
{lixiaoming@ntu.edu.sg, fang.hui@mail.shufe.edu.cn, zhangj@ntu.edu.sg}

## Abstract

We rethink the link prediction problem in signed social networks by also considering "no-relation" as a future status of a node pair, rather than simply distinguishing positive and negative links proposed in the literature. 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. Grounded on the framework, we also design a trinary classification model, and experimental results show that our method outperforms the state-of-the-art approaches.

## Introduction

The increasing interest in signed social networks has brought great impact on many traditional research topics, one of which is link prediction. Link prediction in unsigned networks mainly aims to predict the future connection status between two nodes (linked or not). But, connection status in signed networks can be three situations: positive, negative or no-relation, enlarging the difficulty of link prediction.

On the other hand, link sign prediction in signed networks focuses on predicting signs of existing links (Leskovec, Huttenlocher, and Kleinberg 2010), which is a binary classification problem. In other words, it ignores the no-relation status. However, most link prediction applications in signed networks cannot be simply treated as 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 that ignore the no-relation status are not directly applicable. In addition, 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 identify a relatively new research problem which takes an initial step to consider 'no-relation' as a future dyad status for link prediction in signed networks. Besides, we investigate the link prediction from the temporal dimension. We then propose a structured feature framework for the new problem on the basis of a thorough feature analysis to reveal the underlying mechanism regarding link formation in signed networks. This framework, grounded on both well-known theories and sound observations, can serve as a leading guidance for research on the new problem.



## Problem Statement

Formally, let $G = (V, E^P, E^N, X)$ denote a signed social network, where $V$ is the node set; $E^P$, $E^N$ are the positive and negative link sets respectively; $X$ refers to the no-relation set. $G_t = (V, E^P_t, E^N_t, X_t)$ 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 $(i, j)$ where $x_{ij} \in X_t$, to predict the connection status of $x_{ij}$ at time $t+1$ , which can belong to $E^P_{t+1}$, $E^N_{t+1}$ or $X_{t+1}$.*

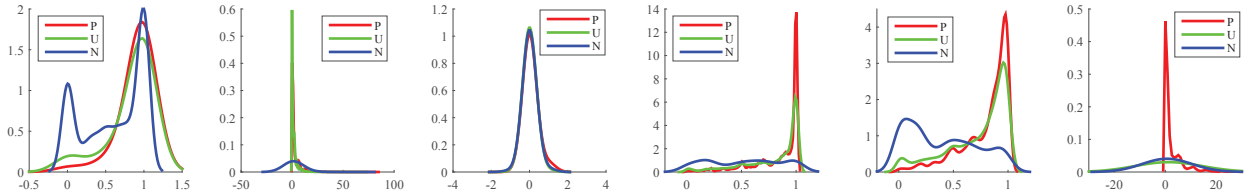
## Feature-based Link Prediction

We first identify and combine the representative features to discriminate the three link statuses, and further propose a feature-based model to minimize the error between the predicted link status and the ground truth:

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

where $S_{ij}$ is the ground truth of link status. L$(\cdot)$ is the link prediction function with the value of 1, -1 or 0, which represents positive, negative or no-relation respectively. In this paper, the multinomial logistic regression model is adopted as link prediction function. Variables mainly consist of two parts: per-dyad side, $u_{ij}$ is the feature set of dyad $(i,j)$; per-node side, $f(u_i, u_j)$ is the function to leverage node-side feature $u_i$ and $u_j$. Loss function $l(\cdot,\cdot)$ is user-specified and application-depended. $\frac{\lambda_1}{2}\|\alpha\|_2^2$ and $\frac{\lambda_2}{2}\|\beta\|_2^2$ are regularizers.

For link prediction in signed networks, 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 the previous studies (Tang, Chang, and Liu 2014) on both unsigned and signed network scenarios, but also derive new features based on our analysis and observations. We then combine and summarize these features into six major categories, and both theoretically and experimentally explore the influence of each category on link formation in signed networks, as follows.




(a) ppratio (b) stadiff (c) recipro (d) poutraito (e) prprs (f) ppff

Figure 1: Kernel smoothed density distribution of selected features: a) ratio of $i \overset{+}{-} w \overset{+}{-} j$ for dyad ($i$,$j$) and their neighbors $\{w\}$; b) status difference between $i$ and $j$; c) status of $j \overset{s}{\rightarrow} i$, where $s = +1$, 0 or -1; d) ratio of positive links sent from $i$; e) multiplication of $i$'s poutratio and $j$'s pintoratio; f) # of $i \overset{+}{\rightarrow} w_i \overset{+}{\rightarrow} j$. P, U and N in legend represent positive, no-relation and negative link respectively. All features pass statistic test except "recipro", which will be dropped.

**Balance Theory** can be simply explained as "my friend's friend is my friend", or "my enemy's enemy is my friend". The logic is to firstly observe the link statues among nodes $i$, $j$ and their common neighbor $w$, and then indicate dyad status ($i$,$j$), which is expected to make the relationship among the three nodes to be more 'balanced'.

**Status Theory** refers to that, a positive link $i \rightarrow j$ means node status of $j$ is higher than $i$. Therefore, we can infer that, if link $i \rightarrow w$ and $w \rightarrow j$ are both positive, link $i \rightarrow j$ is more likely to be positive as status of $j$ is higher than $i$. For each common neighbor of $i$ and $j$, we can calculate the status difference between $i$ and $j$.

**Reciprocity** is the tendency that two nodes with bidirectional links between each other might have the same sign. If we know the link status of $i \rightarrow j$, we can infer the status of the backward link $i \leftarrow j$.

**Rich-get-richer** mainly captures the per-node side features. We consider the features such as positive/negative outdegree of $i$ and positive/negative in-degree of $j$.

**Clustering** adopts the similar insight with link prediction in unsigned networks. It measures per-dyad side features like the number of common neighbors.

**Frequent Subgraph** considers the number of triads and quads constructed by nodes $i$, $j$ and their neighbors.

We extend each category above by involving no-relation status. Take balance theory as an example. It is extended as that no-relation status can make the graph more balanced when $i$ and $j$ have largely the same number of 'friends' and 'enemies'. Meanwhile, in status theory, the dyad ($i$, $j$) tends to have no-relation if their statuses are nearly equal.

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. Figure 1 shows the density distribution of some selected features.

## Experimentation

We conduct experiments with Epinions dataset, which is divided into 3 parts by the timestamp: 'past', 'present' and 'future'. Training set consists of the links formed during 'present' and their corresponding features are measured in the 'past'. Similarly, the testing set is defined as the dyads in the 'future'. We compare the proposed model with existing methods (Song and Meyer 2015), on the basis of AUC (binary classification over 1 and not 1) and GAUC (trinary classification over 1, -1 and 0). As shown in Table 1, our method outperforms others more than 10% for both measurements.

Table 1: Experimental Results

| **Methods** | **GAUC** | **AUC** |
|---|---|---|
| Common Neighbors | 0.576 | 0.587 |
| Katz Measure | 0.591 | 0.592 |
| Singular Value Decomposition | 0.636 | 0.651 |
| Matrix Factorization | 0.654 | 0.645 |
| Optimization-GAUC (2015) | 0.715 | 0.72 |
| **Feature-based Model** | **0.827** | **0.799** |

## Future Work

In the future, we will investigate more real-world datasets to further evaluate the significance of the new problem and the effectiveness of our approach. We will also explore more features and design an advanced model specifically for link prediction in signed networks.

## Acknowledgments

The work is supported by the Telenor-NTU Joint R&D funding awarded to Dr. Jie Zhang.

## References

Leskovec, J.; Huttenlocher, D.; and Kleinberg, J. 2010. Predicting positive and negative links in online social networks. In *WWW*.

Song, D., and Meyer, D. A. 2015. Recommending positive links in signed social networks by optimizing a generalized auc. In *AAAI*.

Tang, J.; Chang, Y.; and Liu, H. 2014. Mining social media with social theories: a survey. *ACM SIGKDD* 15(2):20–29.