LINK PREDICTION AND RECOMMENDATION IN SIGNED SOCIAL NETWORKS

XIAOMING LI

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

20th Jan, 2020	Xiaoming Li
.....	.....
Date	XIAOMING LI

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

20th Jan, 2020

.....

Date

.....

Prof. Jie Zhang

Authorship Attribution Statement

This thesis contains material from 3 papers published in the following peer-reviewed papers accepted at conferences in which I am listed as an author.

Chapter 3 is published as **Xiaoming Li**, Hui Fang, Jie Zhang, "A feature-based approach for the redefined link prediction problem in signed networks," in *Proceedings of the International Conference on Advanced Data Mining and Applications (AMDA)*, 2017, pp. 165–179.

The contributions of the co-authors are as follows:

- I designed the model, conducted data analysis and experiments, and prepared the manuscript drafts.
- Prof Jie Zhang and Prof. Fang Hui provided the initial project direction and revised the manuscript drafts.

Chapter 4 is published as **Xiaoming Li**, Hui Fang, Jie Zhang, "File: A novel framework for predicting social status in signed networks." in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 330–337.

- I designed the model, conducted data analysis and experiments, and prepared the manuscript drafts.
- Prof Jie Zhang and Prof. Fang Hui provided the initial project direction and revised the manuscript drafts.

Chapter 5 is published as **Xiaoming Li**, Hui Fang, Jie Zhang, "Supervised User Ranking in signed networks." in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 184–191.

- I designed the model, conducted data analysis and experiments, and prepared the manuscript drafts.
- Prof Jie Zhang and Prof. Fang Hui provided the initial project direction and revised the manuscript drafts.

20th Jan, 2020

....

Date XIAOMING LI

Acknowledgements

I would like to express my sincere gratitude to my advisor Prof. Jie Zhang for the continuous support of my Ph.D. study, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

My sincere thanks go to my co-advisor Prof. Hui Fang. She kept me going on and this work would not have been possible without her help. I can't thank her enough. I am so grateful for everything she has done for me.

Lastly, I would like to thank my friends and my family for supporting me through the tough days.

Abstract

Link prediction is a fundamental research issue in social networks, which aims to infer the formation of a possible link in the near future. This topic is well studied in the last few years as its significant contributions to improve and enhance online experiences, in the form of further facilitating applications such as recommenders for products or friends, 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.

To fulfill this gap, we first redefine the link prediction problem in signed social networks by also considering "no-relation" as 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, which are positive, negative and no-relation. 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.

Despite the success of the feature-based method, we find that online users are different regarding their activeness and popularity, which actually influence the link formation probability. Besides, "no-relation" status is diverse in social networks. In signed networks, no-relation is the social status apart from positive links and negative links. It is conceivable that most pairs of users with no-relation have limited common connections, however, in reality, many user pairs keep no-relation status even though they have many common connections. It is easy to mispredict

no-relation having many common neighbors as a linked status. Therefore, we take a deep investigation on the diversity of "no-relation" status and 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 to represent users' intrinsic personality, and two explicit features by extending social theories to represent external social influence. We learn these features for each user via matrix factorization with a specially designed ranking-oriented loss function. The effectiveness of our approach is verified by the experiments.

Further, we study the user ranking problem in signed networks, which tries to optimize the link recommendation performance from a personalized perspective. For a certain user, we aim to rank his potential "friends" on the top whereas rank "enemies" on the bottom. Current approaches focus on global ranking thus cannot provide effective personalized ranking results. Besides, they have a relatively unrealistic assumption that each user treats her neighbors' social strengths indifferently. In this work, we propose a supervised method based on random walk to learn social strengths between each user and her neighbors, in which the random walk more likely visits "potential friends" and less likely visits "potential enemies". We learn the personalized social strengths by optimizing on a particularly designed loss function oriented on ranking. Experimental results demonstrate the superiority of our approach over the state-of-the-art approaches.

To sum up, we have proposed a series of approaches for link prediction in the signed network scenario. These approaches are constructed in a more realistic setting and can be used in real-world applications.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Online social networks (OSN) have become an essential part of modern social life [1]. Popular platforms such as Facebook, Twitter, and Instagram allow users to follow or add friends with each other online. If we present a social network as a graph, users will be the nodes and social interactions/connections are the links in the graph. The link prediction problem thus tries to infer the formation of possible social connections in the near future. It's fundamental research in the social network domain and has attracted attention both in academia and industry [2]. Numerous works have demonstrated that link prediction can improve and enhance online experiences [3, 4], in the form of further facilitating applications such as recommenders for products or friends [5] and social networks [6].

Nowadays, more OSN platforms are constructed based on the signed structure. 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 have broad implications for real businesses nowadays, and online systems such as Slashdot[1], Epinions[2] and Wikipedia voting community and Bitcoin exchange community. Here we briefly introduce those signed networks used in this thesis. Epinions is a product review website, where users can specify their trust or distrust toward other users. Therefore, links between users can be positive or negative, which indicates the trust and distrust relationships. Slashdot is a social news website focusing on tech-community interactions and discussion. In Slashdot, users can add others as friends or foe, which is represented as positive links (trust) and negative links (distrust) in the network topology. Wikipedia allows users to vote for others to the role of admin. A positive link represents a supporting vote and a negative link represents an opposing vote. Bitcoin dataset are extracted from user interactions in an exchange, where users can rate others as trust or distrust. Besides, 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, the relationship between online users is not only limited to positive (e.g. friend and trust) anymore, but tries to add more alternatives to be consistent with the human relationship in real life.

The increasing interest in signed social networks has brought a great impact on many traditional research topics, one of which is link prediction. 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, the connection status of two nodes in signed networks could be positive, negative and no-relation, which increases the difficulty of link prediction. Figure 1.1 illustrates the link prediction problem in signed social networks. Given a snapshot of the current network, we

[1]www.slashdot.org
[2]www.epinions.com

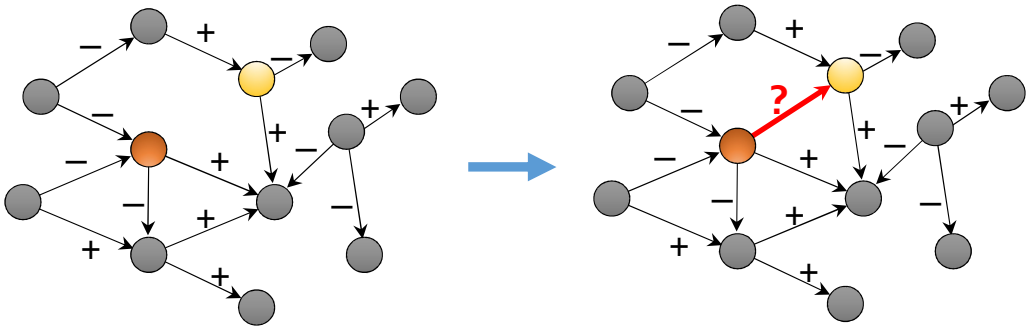


FIGURE 1.1: Illustration of link prediction in signed social networks.

aim to predict the social status of the unconnected user pair in the future, which can be positive, no-relation, or negative.

1.2 Research Gaps

Most studies [7–10] 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 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. In other words, this kind of research basically ignores 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. 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. [11] 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.

In fact, 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 about 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.

Second, most attempts [7, 12–14] are based on classification rather than ranking. They aim to distinguish positive and negative links. We argue that a more realistic setting should be ranking these user pairs by the order of positive, no-relation and negative status. Besides, current approaches do not take attention to the diversities of no-relation status. 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[3], $40,779$ 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.

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

The third research gap is the link recommendation problem in signed networks, or personalized user ranking problem. It aims to provide a user ranking list by the order of link formation probability for a certain user. In unsigned networks, this topic has been well studied, which includes similarity-based ones [15, 16], random walk based models [17, 18] and low-rank models [19–21]. However, these existing methods cannot be directly applied in signed networks because of the existence of negative links. Therefore, a few works have strived to extend the traditional methods into the signed scenarios, like Similarity-based approaches [22, 23] and random walk approaches [24, 25]. However, these studies are mainly extended models in unsigned scenarios and they focus on global ranking rather than the personalized perspective. Besides, they assume all the links in the network have the same weights. In other words, they ignore the difference in social strengths, which actually play a key role in personalized user ranking.

To summarize, three research gaps are as follows: first, no-relation status is ignore in current link prediction approaches; second, current studies are mainly based on classification rather than ranking; third, the task of personalized ranking in not investigated in signed scenarios.

In this thesis, we aim to answer three research questions.

- RQ1: What is the underlying mechanism regarding link formation in signed networks?
- RQ2: Given two user pairs in signed networks, which pair is more likely to become friends (or enemies)?
- RQ3: Given a seed user in signed networks, how to rank other users by the probability of link formation with her?

1.3 Research Approaches

To address the proposed research questions, three pieces of works have been proposed in this thesis. They are summarized as follows:

A feature-based approach to understand link formation mechanism in signed social networks [26]: To answer RQ1, 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 an 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 [11].

A novel framework of integrating latent and explicit features [27]: To answer RQ2, 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 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.

A supervised user ranking model based on random walk [28]. To answer RQ3, we propose Signed Supervised Random Walk (SSRW), through which we learn social strengths that capture a user's different preferences towards different neighbors, and thus to better facilitate the task of personalized user ranking. Current approaches aim to generate a global ranking list for the whole network, which could easily lead to a relatively unfair scenario where some users might have a large number of potential links in the ranking list while most users have very few or even no links. In this case, they cannot be easily adapted for many real-world applications such as social recommendation or social-aware product recommendation. In contrast, personalized user ranking, which generates a ranking list for each individual, is more practical and realistic [29]. Besides, the ranking list for a user provided by existing random walk methods is fixed given a certain network snapshot (i.e., the network structure). They inappropriately assume all the links have the same weights (i.e. social strengths, a.k.a. link strengths). In other words, they cannot learn each individual's own opinions towards her neighbors, such as what kind of user link (i.e. neighbors) is more important. In this work, instead of

considering the random walk in a given network snapshot (i.e., training data), we split the training data into two parts in terms of the timestamp (denoted as A and B), and learn social strengths (i.e., transition probabilities) so that random walk more likely visits those newly positively connected nodes (i.e., in B compared to A) whereas more reluctantly visits the newly negatively connected nodes. We conduct experiments on four real-world datasets and the results show that SSRW's performance has an improvement of 6.05% compared to the state-of-the-art approaches. To improve SSRW's efficiency but simultaneously maintain its effectiveness, we also design a fast ranking method (F-SSRW) based on the local structure among each seed node and a certain set of candidates. It has been demonstrated that F-SSRW can maintain the performance in contrast with the original SSRW when the ranking candidates of a user satisfy the requirement of having substantial common neighbors with the user.

These 3 works aim to solve different research questions. We first aim to understand link formation mechanism by feature-based study. Then we use FILE model to rank user pairs based on their link formation probability, which is designed from the view of the social network platform. The third work aims to optimize personalized ranking, which is designed from the view of individual users.

Our proposed approaches can be adopted in many real-world applications. For example, positive link prediction can be used for friend recommendation. By considering negative link, our approaches can be used in more areas especially in the security domain. If we treat the social interaction as a link, negative link prediction can be applied for social network spam detection, email spam detection, P2P-based collaborative spam detection, etc. Another security-related application is voting investigation and prediction. Other applications include anomaly detection, criminal detection and tracking terrorist networks.

1.4 Contributions

The contributions of this thesis can be summarized as follows:

- 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 of any two nodes, rather than distinguishing the sign of a certain link, which is the common setting of the current approaches. Our setting is more realistic and can be served as a guidance and elicit more related research in this area.

- We propose a structured feature framework on the basis of a thorough feature analysis to reveal the underlying mechanism regarding link formation in signed networks. We not only adopt existing features in the previous studies [7, 30] on both unsigned and signed network scenarios, but also derive new features based on social theories and observations. The feature framework fills the research gap as the first feature engineering study considering the three social statuses together.

- 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 social networks. We also find that the link prediction problem in signed networks becomes rather difficult mainly due to the diversity of no-relation, as it is easy to mispredict no-relation having many common neighbors as a linked status. Therefore, we propose the FILE framework which considers both the social linkage criteria of individual users and the external social influence from the neighborhood of every user pair.

- We propose a novel link prediction framework that integrates social explicit features into a latent model. We demonstrate that this can significantly improve positive link, negative link, and no-relation prediction.

- We propose a supervised ranking model SSRW which learns social strengths to optimize the user ranking in signed networks. Based on the heuristics from data analysis, we further design a simplified and efficient ranking method (F-SSRW), which only focuses on certain candidate nodes and runs the learning algorithm within the local graph of the seed node. A comprehensive evaluation demonstrates the superiority of the proposed models over state-of-the-art approaches, and the robustness in terms of parameters and experimental settings.

1.5 Thesis Organization

This thesis is organized as follows:

- Chapter 2 carries out a comprehensive survey on related works in this area, including link prediction approaches in unsigned social networks, sign prediction models and works on personalized user ranking.

- Chapter 3 proposes a feature framework on the basis of a thorough exploration of potential features.

- Chapter 4 introduces the model which integrates both Latent and Explicit features (FILE), to better deal with the no-relation status and improve the overall link prediction performance in signed networks.

- Chapter 5 proposes a personalized user ranking method, which is based on random walk to learn social strengths between each user and her neighbors,

and make the random walk more likely visits 'potential friends' and less likely visit 'potential enemies'.

- Chapter 6 draws a conclusion for the thesis and point out the directions for future works.

Chapter 2

Related Work

In this chapter, we review the state-of-the-art approaches which are related to our study. First, we give detailed reviews of approaches for link prediction in unsigned networks. Then, we survey the latest works on the link signed prediction. Finally, we provide reviews on personalized user ranking methods.

2.1 Link Prediction in Unsigned Networks

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 the near future [3]. Links in traditional social networks do not have signs, therefore, link prediction in unsigned networks considers only two possible future connection statuses of two nodes, i.e., linked or not linked. Generally, link prediction methods can be divided into two classes: unsupervised and supervised methods.

2.1.1 Unsupervised Approaches

Unsupervised methods mainly try to define a score function for node pairs. The fundamental assumption of these approaches is that users tend to form links with other similar users. Therefore, the core task is to define a similarity metric with a reasonable heuristic assumption. Unsupervised methods consist of local neighbor based metrics and path-based metrics [3].

Popular neighbor-based metrics include: the number of common neighbors [31], Adamic/Adar Index [31], Jaccard Coefficient [3], Preferential attachment [32], Resource Allocation Index [33], Resource allocation based on common neighbor interactions[34],Average Commute time [35], The Hub Promoted Index [36], The individual Attraction Index [37], Local Leicht-Holme-Newman Index [38], Mutual Information [39], Local Naive Bayes [40], Car based Indices [41] and Functional Similarity Weight [42]. These metrics are derived from the local neighborhood structure of the user pair. They are easy to implement and have a good performance against complicate models. However, they only adopt limited structural information of users whereas the information of the whole network topology is ignored.

Meanwhile, the features related to the path between two nodes in a network structure are also used to compute the similarities of node pairs. Compare to neighbor based methods, path-based methods consider more topological information to define the score function. Typical approaches include: Katz [43], Vertex Collocation Profile [44], Random Walk with Restart [45],The Blondel Index [46], ProfFlow [47], SimRank [48], Pseudoinverse of the Laplacian Matrix [49], Random Forest Kernel Index [50], FriendLink [51]. Path-based approaches always have high complexities but their performance is relatively better than neighbor-based ones since they consider more information.

2.1.2 Supervised Approaches

Popular supervised methods include feature-based classification models and latent feature models. link prediction in unsigned networks can be treated as a classification problem as there are two types of social status: linked or not-linked. Therefore, several classification models [52] can be adopted for the classification. Hasan et al. [52] identified topological features and proximity features, and showed that most well-known classification models can perform well in link prediction, which include SVM, K-NN, Decision Trees, Naive Bayers, Bagging, .etc.

Meanwhile, latent models [53, 54] learn the latent vectors of each user and define a link score function based on users' latent vector. They have shown a good prediction performance. The latest works start to apply deep learning techniques. In [55], authors learn node embeddings via graph attention, and make predictions based on the node embedding. Gu et al [56] propose Deeplinker to extract vertex representation, and they use link information as features rather than to learn node representation. Wang et al [57] propose a model based on deep convolutional neural network which can capture the features from the sub-graph.

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.

2.2 Sign Prediction in Social Networks

The approaches discussed in Section 2.1 require all links in the network to be positive. These models can not directly be extended to signed social networks [58].

Current approaches mainly treat link prediction problems as a binary classification problems, to distinguish positive links and negative links. We summarize these works as sign prediction problem for clarification.

2.2.1 Feature-based Approaches

As indicated, existing attempts 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. [7] 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. They showed that more than 90% of user triads satisfy balance theory and status theory, and the prediction accuracy is more than 90% in Epinions and Slashdot dataset.

In order to construct more features, k-cycle-based features are proposed in [8] where triangle-based features ($k = 3$) are especially 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. [12] leverage the pattern of frequent subgraph among node pairs, to predict link status. Patidar et al. [13] adopt users' attributes as features, and train a classifier to infer the sign. Dubois et al. [14] adopt users' attributes as features and interaction features.

2.2.2 Latent Model Approaches

Another type of popular method is low-rank model. For example, Hsieh et al. [9] verify that signed networks naturally present a low-rank structure, and a matrix factorization model is proposed to infer link signs. Agrawal et al. [59] also follows the same idea and adopt pairwise empirical error as the loss function. For dynamic

networks, Cen et al. proposed a low-rank tensor model [60] and learn the model via dictionary learning.

However, all the aforementioned methods assume the existence of links with uncertain signs. Song and Meyer [11] also adopt a low-rank model to infer link signs, which learns the latent features by minimizing the generalized AUC loss. The major purpose of the work is to distinguish positive and negative link status, and the no-relation information is used in the training period of the model.

Hsieh et al. [9] 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 [61] 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. [62] 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 a non-zero value.

2.3 Personalized User Ranking in Signed Networks

Different from the traditional link prediction, the personalized user ranking in signed networks tries to provide a ranking list for each individual user, by the order of "potential friends" to "potential enemies". The result can be directly used

in recommendation applications. It is worth noting that there are lots of existing works for personalized user ranking in unsigned networks, and the representative approaches include similarity-based approaches [15, 16], random walk based models [17, 18] and low-rank models [19–21]. However, these existing methods cannot be directly applied in signed networks because of the existence of negative links. Therefore, a few works have strived to extend the traditional methods into the signed scenarios, which also can be summarized into two categories: similarity-based approaches and the random walk based approaches.

Similarity-based approaches try to define a link score metric to measure the similarity in signed social networks. The key issue is to define the similarity based on a reasonable assumption. Symeonidis et al. [22] propose a similarity metric based on users' out/in degree of positive and negative links. A higher similarity score between two users indicates a higher chance to establish a positive link, while a lower score indicates a possible negative link. Zhu et al. [23] use the number of common friends minus the number of common enemies as the similarity metric. Song et al. [63] learn users' latent vectors by adopting matrix factorization technique, and model the ranking score as the inner product between the corresponding user vectors.

For random walk based approaches, Shahriari et al. [24] firstly split the signed graph into two graphs: a positive and a negative one, and then apply random walk with restart on each graph. They finally combine the results from two random walks to generate one ranking list for each user. In [25], a signed network is converted into a positively weighted graph, and then obtain the ranking list using the random walk technique. Jung et al. [29] propose a model named SRWR, which introduces a sign into a random surfer so that negative links can be also considered by changing the sign of walking.

In summary, these studies are heuristically extended from the methods in unsigned scenario. Meanwhile, current approaches mainly focus on global ranking rather than the personalized perspective. Besides, they assume all the links in the network have the same weights. In other words, they ignore the difference in social strengths, which actually play a key role in personalized user ranking.

2.4 Summary

This section provides a systematic summary of the related research. We first briefly discuss traditional link prediction approaches in unsigned networks. Then we make a comprehensive survey on sign prediction in social networks, which include feature-based approaches and latent model approaches. Last, we show the state-of-the-art studies on personalized link recommendation in signed networks. We also briefly discuss the strength and weakness of current works.

Existing works treat link prediction in signed networks as a sign prediction problem, with the assumption of link existence. 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. We need to re-investigate the features adopted in the literature for link prediction and design more specific approaches for the newly redefined problem (i.e., link prediction in signed networks) on the basis of a better feature design and problem analysis. Besides, current approaches mainly focus on global link prediction rather than the personalized perspective. Besides, they assume all the links in the network have the same weights. In other words, they ignore the difference in social strengths, which actually play a key role in personalized user ranking. To tackle this issue, we propose a series of link prediction models.

Chapter 3

A Feature-based Approach for Link Prediction in Signed Networks[1]

As discussed in Section 2.2, most link prediction applications in signed networks cannot be simply treated as the sign prediction problem. 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.

This chapter proposes 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.

[1]The work in this chapter has been published as [26] in ADMA 2017.

TABLE 3.1: Our research question compared to related works.

Link Prediction (LP) Scenario	**Classification Target**	**Features**	**Theory supports**
Link Prediction in Unsigned Network	Linked/ Not-linked	CN,Adamic/Adar, JC, Katz distance RAI	Similarity Closeness
Link Prediction in Signed Network	Positive/ Negative	Triad based features topological features	Balance theory Status theory
Redefined Link Prediction in Signed Networks	Positive/ Negative/ No-relation	Features derived from Social Theories	Balance Theory Status Theory Reciprocity Rich get richer Clustering Frequent Subgraph

Table 3.1 shows the difference with previous works. In this work, we first thoroughly explore features derived from social theories 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 [11].

The main contributions of this chapter are two-fold:

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 [7–9].

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.

The chapter is organized as follows: we first give the formal statement of the redefined link prediction problem in signed social networks in Section 3.1. In section 3.2, we present the feature framework and introduce the features derived from social theories. In Section 3.3, we proposed our feature-based approach. we also discuss the feature selection mechanism and the method to handle the imbalance issue. Experimental results are shown in Section 3.4. The notations used in this chapter are summarized in Table 3.2.

TABLE 3.2: Notations.

G	signed social network
V	the node set in G
E^P	the set of positive links in G
E^N	the set of negative links in G
X	the set of no-relation in G
G_t	the snapshot of the network at time t
(i, j)	social status between node i to j
S_{ij}	the link sign from user i toward user j
O_i	user i's outgoing link set
I_i	user i's incoming link set
C_{ij}	the set of the common neighbors between users i and j

3.1 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 $x_{ij} \in X_t$, 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 chapter, 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 do we 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. We also introduce two techniques to address the data imbalance issue for link prediction in signed networks.

3.2 Feature Framework

3.2.1 Data Description

We first introduce the datasets used in this work. We obtain four publicly available datasets[2] with the signed structure, i.e., Epinions,Slashdot, Wikipeida and Bitcoin. In both social networks, users can establish trust and distrust relationship, i.e., positive or negative links with other users. Table 3.3 provides descriptive statistics for these datasets.

TABLE 3.3: Descriptive statistics of users and links in the Datasets

	Epinions	Slashdot	Wikipedia	Slashdot
Users	131,828	82,140	9,654	3,783
User pairs with positive links	717,667	425,072	87,766	22,650
User pairs with negative links	123,705	124,130	16,788	1,536
User pairs with no-relations	$1.7 * 10^{10}$	$6.7 * 10^{9}$	$9.3 * 10^{7}$	$1.4 * 10^{7}$

From Table 3.3, we can quickly summarize two general data patterns that occur in signed networks: (1) sparsity: a signed network is quite sparse, and we find that there are no more than 10.2% users with degree ≥ 25 in all 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.

3.2.2 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

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

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 [7, 30] 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 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.

3.2.3 Feature Definition

Balance Theory [64] 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 users 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 [7, 65] 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 indicates 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 [66] 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 [67] indicates two active or popular users will more likely get linked. Thus, we derive 10 features to capture this phenomenon. Given two users 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 [3] 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** [3], **Katz** [43], **JC** [3], **PA** [32], **Status Similarity** [22]. In signed network, a smaller feature in this category indicates a higher chance to have no-relation.

Frequent Subgraph [7, 8] 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 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 Figure 3.1.

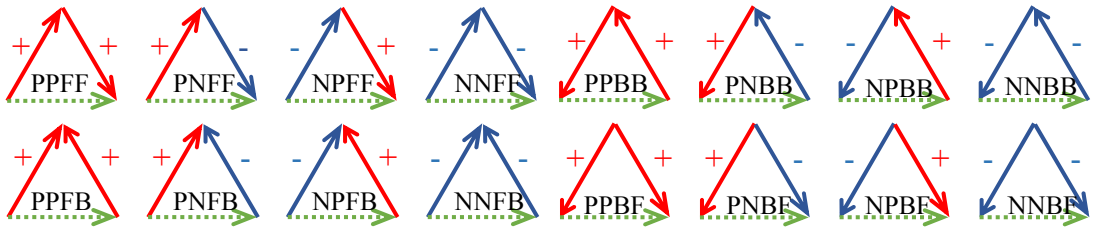


FIGURE 3.1: The sixteen triads are fundamental and crucial units for network topology analysis.

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.

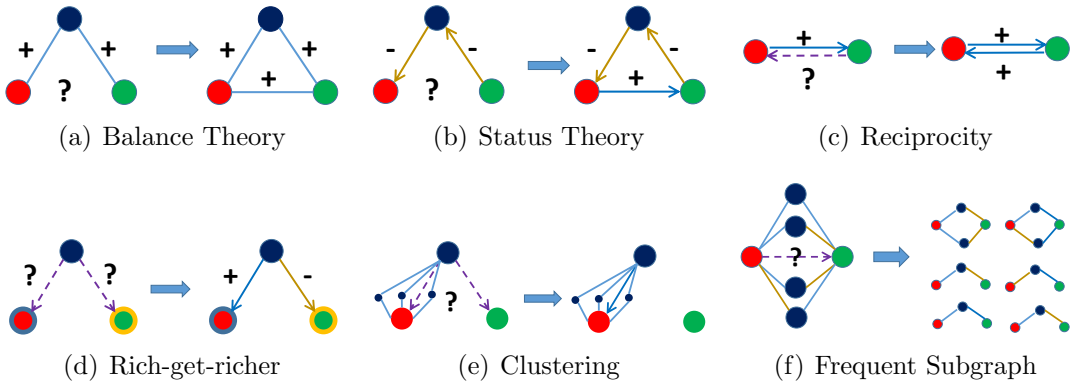


FIGURE 3.2: Illustration of six social theories.

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 3.4 summarizes the full list of the features we derived from social theories.

3.3 Feature-based Approach

The proposed feature-based model can be stated as:

$$\min_{\alpha,\beta} \sum l(S_{ij}, L(\alpha f(u_i, u_j) + \beta u_{ij})) + \frac{\lambda_1}{2}\|\alpha\|_2^2 + \frac{\lambda_2}{2}\|\beta\|_2^2 \tag{3.1}$$

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**.

3.3.1 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 first 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

TABLE 3.4: Features derived from social theories

Feature	Notation
Balance Theory	
pp	ppff + ppfb + ppbf + ppbb
nn	nnff + nnfb + nnbf + nnbb
pn	$\|C\|$− pp − nn
pp_ratio	pp/$\|C\|$
nn_ratio	nn/$\|C\|$
bal_diff	pp + nn − pn
Status Theory	
sta_diff	sta_diff_p − sta_diff_n
sta_diff_p	ppbb + nnff + pnfb + npbb
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	$\|O_i^+\|$
out_n	$\|O_i^-\|$
in_p	$\|I_j^+\|$
in_n	$\|I_j^-\|$
out_p_ratio	$\|O_i^+\|/\|O_i\|$
in_p_ratio	$\|I_j^+\|/\|I_j\|$
prprs	$(\|O_i^+\|/\|O_i\|) * (\|I_j^+\|/\|I_j\|)$
prnrs	$(\|O_i^+\|/\|O_i\|) * (\|I_j^-\|/\|I_j\|)$
nrnrs	$(\|O_i^-\|/\|O_i\|) * (\|I_j^-\|/\|I_j\|)$
nrprs	$(\|O_i^-\|/\|O_i\|) * (\|I_j^+\|/\|I_j\|)$
Clustering	
cn	$\|C\|$
Katz	ppff + pnff + npff + nnff
Jaccard coefficient	$\|C_{ij}\|/(\|O_i\| + \|O_j\| + \|I_i\| + \|I_j\|)$
preferential attachment	$(\|O_i\| + \|I_i\|) * (\|O_j\| + \|I_j\|)$
status similarity	$1/(\delta(i) + \delta(j) - 1)$, $\delta(i) = \|I_i^+\| + \|O_i^-\| - \|O_i^+\| - \|I_i^-\|$

those three classes. Figure 3.3 shows the kernel smoothed density distribution of some selected features. As shown in Figure 3.3, 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.

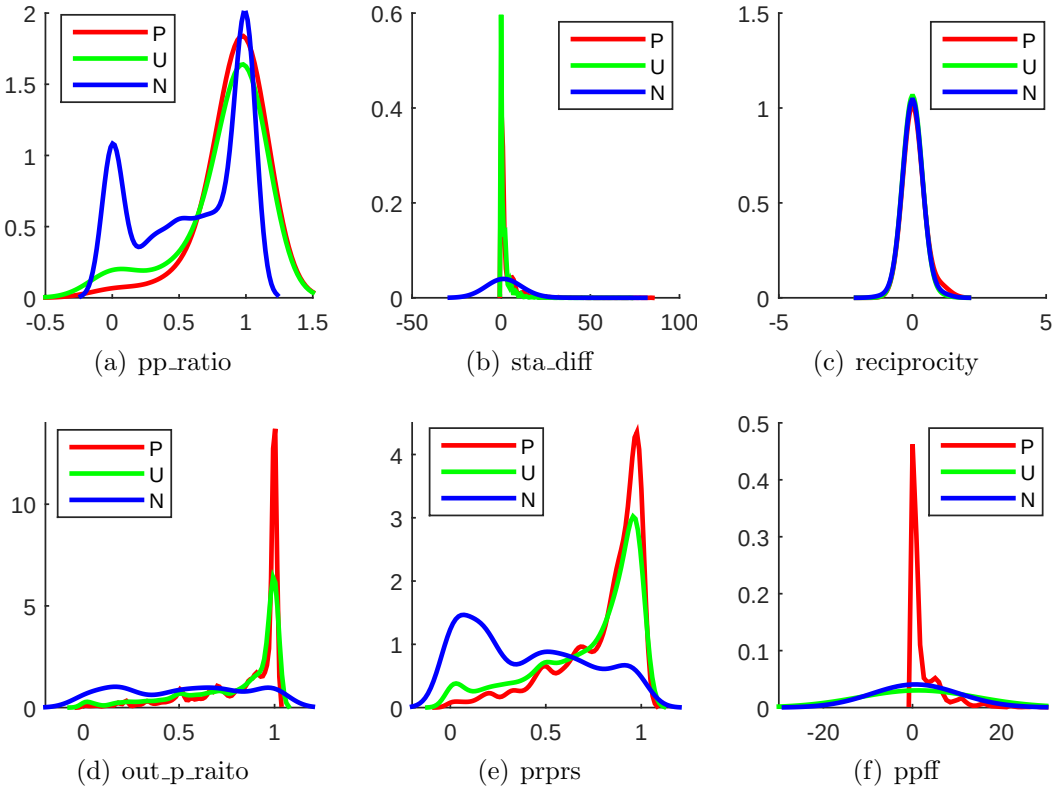


FIGURE 3.3: Kernel smoothed density distribution of selected features.

3.3.2 Handling the 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. 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, 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), 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.

3.4 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.

Generally, we aim to answer two questions:

- Can our proposed approach predict social relations including positive, negative and no-relation?
- Does the introduction of no-relation help to improve the performance of link recommendation?

3.4.1 Experimental Setting

First, we try 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 real-world scenarios. Based on the undersampling method discussed in Section 3.3.2, 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 $T3$ is $18,489$ and $15,741$ respectively.

Since there is no timestamp information in other 3 datasets, 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. In Epinions and Slashdot dataset, we filter user pairs by different number of common neighbors, i.e., minimum as 1, 10

and 25. For other two small datasets, we show the result on user pairs with at least 1 neighbor. 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.

3.4.1.1 Evaluation Metrics

As discussed in Section 3.3.2, we will use measurements of ranking rather than of accuracy. We adopt the generalized AUC (i.e., GAUC) metric, which is defined as:

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

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.

3.4.1.2 Benchmark Approaches

We compare our approach with the following methods:

• Common Neighbors (CN) [3], ranks user pairs by the number of their common neighbors, including both the positive and negative common neighbors.

• Katz [43], 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) [7], adopts a total of 23 features based on topology triads and user degree, and use regression as the learning model.
• Matrix Factorization (MF) [9], 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) [11], 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.

CN and Katz are metric-ranking based methods, All23 is a feature-based supervised learning method, MF and OptGAUC are low-rank models. These 3 types of methods are state-of-the-art approaches and widely used in link prediction problem. In this section, we make a comprehensive comparison between our approach and these methods.

3.4.2 Prediction Performance

As shown in Table 3.5, we can see that our approach outperforms others on both GAUC and PAUC metrics, under different dataset settings. We also run experiments on two small datasets Wikipedia and Bitcoin, and the experimental results are shown in Figure 3.4.

TABLE 3.5: Performance comparison

Method	Epinions						Slashdot					
	cn≥ 1		cn≥ 10		cn≥ 25		cn≥ 1		cn≥ 10		cn≥ 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**

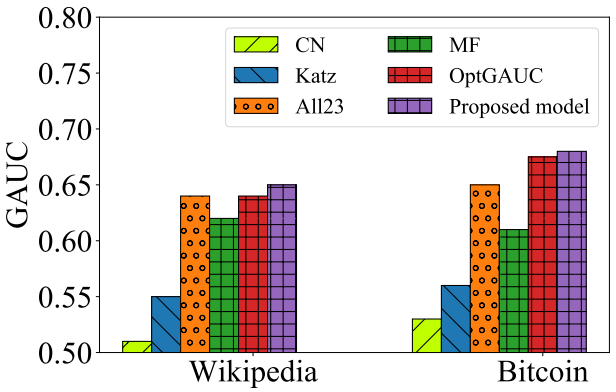


FIGURE 3.4: Experiment results on Wikipedia and Bitcoin

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.

We also examine the performance of different multi-class 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 3.6 indicate that our approach with the multinomial logistic regression model achieves the best performance in terms of GAUC.

3.4.3 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

TABLE 3.6: 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**

category. As shown in Table 3.7, 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 on Slashdot dataset. The full model which adopts all features shows the best performance, demonstrating the effectiveness of feature combinations.

TABLE 3.7: 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 3.8, 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 3.8: 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**

3.5 Summary

In this chapter, 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.

Chapter 4

FILE: A Novel Framework for Predicting Social Status in Signed Networks[1]

In the previous chapter, we propose a feature based approach for the link prediction problem in signed networks. It carries 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 linkage criteria [68], which is not realistic. For example, some users might be more willingly to connect to others while some users are more influential [68] and easily connected by others.

In fact, the link prediction problem in signed networks becomes rather difficult mainly due to the diversity of no-relation. It is conceivable that most pairs of users

[1]The work in this chapter has been published as [27] in AAAI 2018.

with no-relation have limited common connections (*Stranger*), however, in reality, many user pairs keep no-relation status even though they have many common connections (*Frenemy*). For example, in the Epinions dataset, 40779 out of 94732 user pairs who share more than 100 common neighbors still have no-relation with each other. It is easy to mispredict no-relation having many common neighbors as a linked status.

In this chapter, we propose a novel Framework of Integrating 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 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 evaluations demonstrate the effectiveness of the proposed framework on link prediction in signed networks.

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 positive link, negative link and no-relation prediction.

- 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.

The chapter is organized as follows: we first discuss the data analysis results in Section 4.1. In section 4.2, we present the FILE framework and give the optimization method. Experimental results are shown in Section 4.3. The notations used in the chapter are summarized in Table 4.1.

TABLE 4.1: Notations.

S	signed adjacency matrix, $S_{ij} \in \{1, 0, -1\}$
$\|P\|$	the number of positive links in S
$\|U\|$	the number of no-relation in S
$\|N\|$	the number of negative links in S
(i, j)	social status between node i to j
CN	the number of common neighbors between a user pair
u_i	user i's latent vector in term of activeness
v_i	user i's latent vector in term of popularity
f	explicit features
$N(\cdot)$	normalized function
$I(\cdot)$	the 0/1 indicator function

4.1 Preliminaries

4.1.1 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 previous

studies, which aims to classify 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.

4.1.2 Data Analysis

Previous analysis on data patterns in signed networks [69, 70] 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], Wikipedia RFA[3] and Bitcoins[4].

4.1.2.1 Data imbalance.

From Table 4.2, 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 in 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 accuracy metrics for reasonable evaluations and comparisons of different approaches in our experiments.

[2]snap.stanford.edu/data/soc-Slashdot0902.html
[3]snap.stanford.edu/data/wiki-RfA.html
[4]cs.umd.edu/~srijan/wsn

TABLE 4.2: 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$\leq$ 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

4.1.2.2 Stranger v.s. Frenemy.

The statistics in Table 4.2 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 4.1 shows the changing ratio of the no-relation user pairs after 15 months. The 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 a 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.

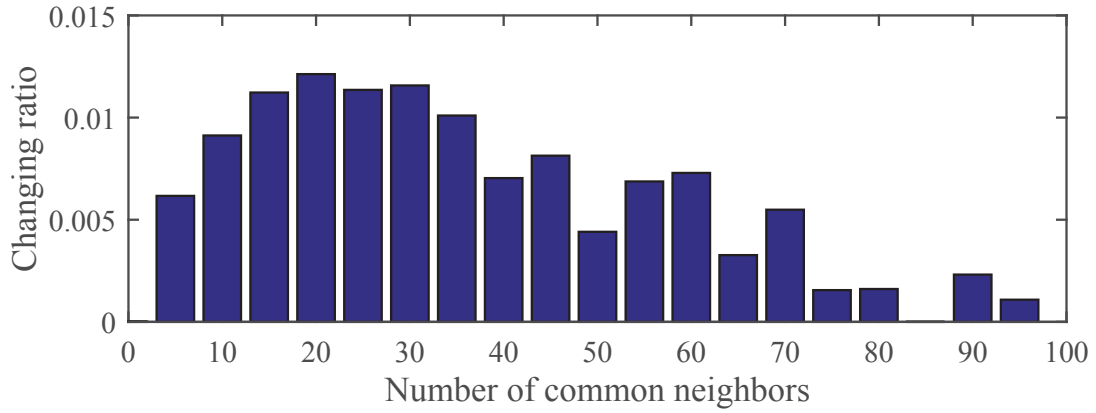


FIGURE 4.1: The distribution of no-relation changing ratio.

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

In view of psychosocial theories, both intrinsic personality [71] and external influence from mutual neighbors [72] 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 4.2 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.

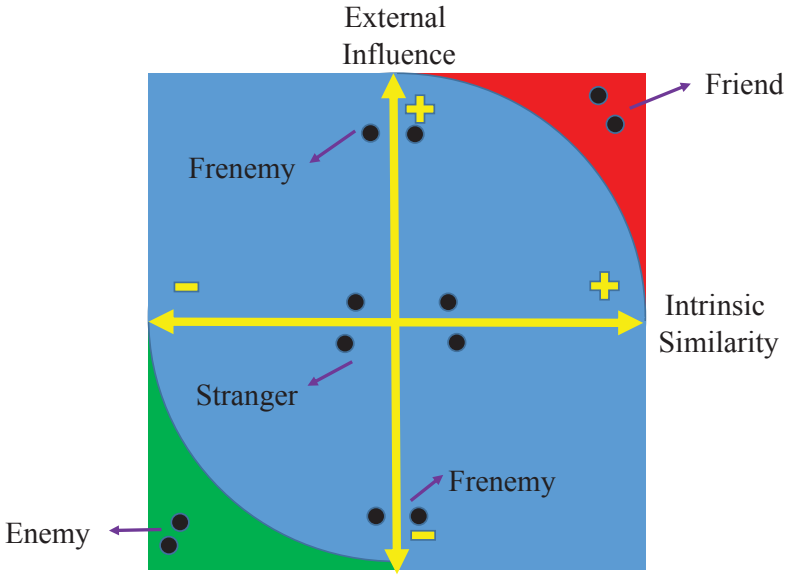


FIGURE 4.2: Illustration of the influential social components.

4.2 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.

4.2.1 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 [9], 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.

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:

$$\mathcal{L}^1(i, j) = u_i^T \cdot v_j \tag{4.1}$$

4.2.2 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 [69, 73] 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 3.1, 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 [69, 73]. Our experimental results show that with only the above two explicit features, our approach can already achieve better results than other existing approaches.

4.2.3 Link Score Function

The link score function is defined as follows:

$$\mathcal{L}(i,j) = \overbrace{N(u_i^T \cdot v_j)}^{\text{Latent}} + \overbrace{\sum_k \alpha_w * N(f_{ij}^w)}^{\text{Explicit}},\ 0 < \alpha_w < 1 \tag{4.2}$$

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 either positive or negative link.

In Equation 4.2, 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.

4.2.3.1 Normalization Function.

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

$$N(x|\theta) = \frac{1 - \exp(-\theta x)}{1 + \exp(-\theta x)} \tag{4.3}$$

[5]Note that as indicated in the explicit features part, more explicit features can be designed and incorporated into Equation 4.2.

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.

4.2.4 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 4.2, the loss function is defined as:

$$\min \sum_{S_{ij}=1} I(L_{ij} \geq 1) + \sum_{S_{ij}=0} I(|L_{ij}| < 1) + \sum_{S_{ij}=-1} I(L_{ij} \leq -1) \tag{4.4}$$

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 4.2, 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:

$$\min \sum_{S_{ij}=1} (1 - L_{ij}) + \sum_{S_{ij}=0} (L_{ij}^2 - 1) + \sum_{S_{ij}=-1} (L_{ij} + 1) \tag{4.5}$$

To construct the equivalent reduced form for Equation 4.5 and add regularizers to avoid overfitting, the loss function F can be rewritten as follows:

$$\min_{U,V} \frac{1}{2} \sum_i \sum_j (1 - S^2)L^2 - SL + \frac{\lambda_1}{2}\|U\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2 \tag{4.6}$$

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:

$$\frac{\partial F}{\partial u_i} = \sum_j \left((1 - S^2)\Delta_1\Delta_2 - S\Delta_2\right) * v_j + \lambda_1 u_i \tag{4.7}$$

$$\frac{\partial F}{\partial v_j} = \sum_i \left((1 - S^2)\Delta_1\Delta_2 - S\Delta_2\right) * u_i + \lambda_2 v_j \tag{4.8}$$

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. 4.7;
  $V_{t+1} = V_t - \beta \frac{\partial F}{\partial V_t}$ based on Equation. 4.8;
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.

4.3 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.

4.3.1 Experimental Setting

As shown in Table 4.3, 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 [74] and an online user has about 338 friends on average [75]. Therefore, it is more realistic to check users with a high degree. This sampling strategy is widely adopted in the previous studies [3]; 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 4.3 where we use 'name@degree' to represent a specific dataset, e.g., Epinions@10 (or E@10) is the dataset about Epinions with $d \geq 10$.

4.3.1.1 Evaluation Metrics.

We use the standard 5-fold cross-validation for training and testing, and utilize GAUC (Generalized AUC over $+1$, 0 and -1) to measure the overall ranking

TABLE 4.3: 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

performance, formulated as:

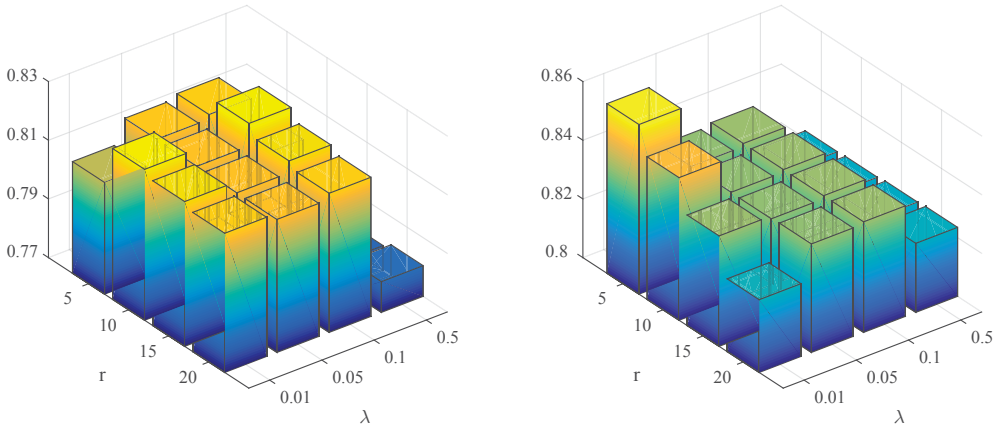
$$\frac{1}{|P|+|N|}\left(\frac{1}{|U|+|N|}\sum_{a_i\in P}\sum_{a_s\in U\cup N} I\left(L(a_i)>L(a_s)\right)+\right.$$
$$\left.\frac{1}{|U|+|P|}\sum_{a_j\in N}\sum_{a_t\in U\cup P} I\left(L(a_j)<L(a_t)\right)\right)$$

where $L(\cdot)$ is the link score function.

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.

TABLE 4.4: 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 4.3: Performance as parameters change w.r.t. GAUC.

4.3.1.2 Benchmarking Approaches.

We conduct comparisons with five state-of-the-art approaches, including feature-based models: Common Neighbors (CN) [3] and Social Feature Model (SFM) [26]; latent models: Low Rank Modeling (LRM) [9] and ranking based latent models of Bayesian Personalized Ranking (BPRMF) [76] and Optimizing GAUC (OptGAUC) [11].

4.3.1.3 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

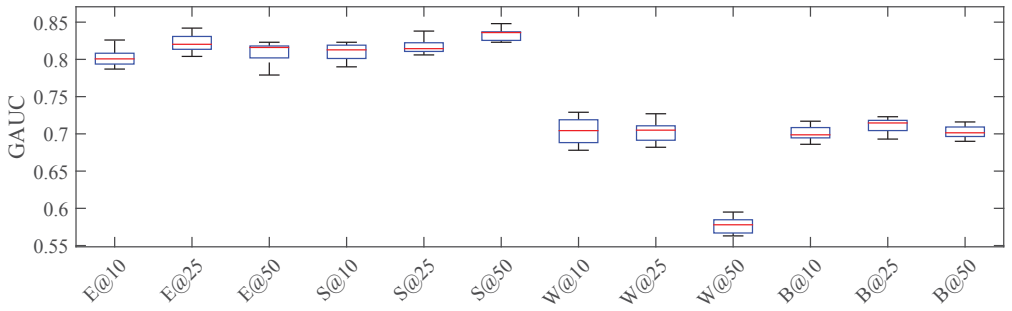


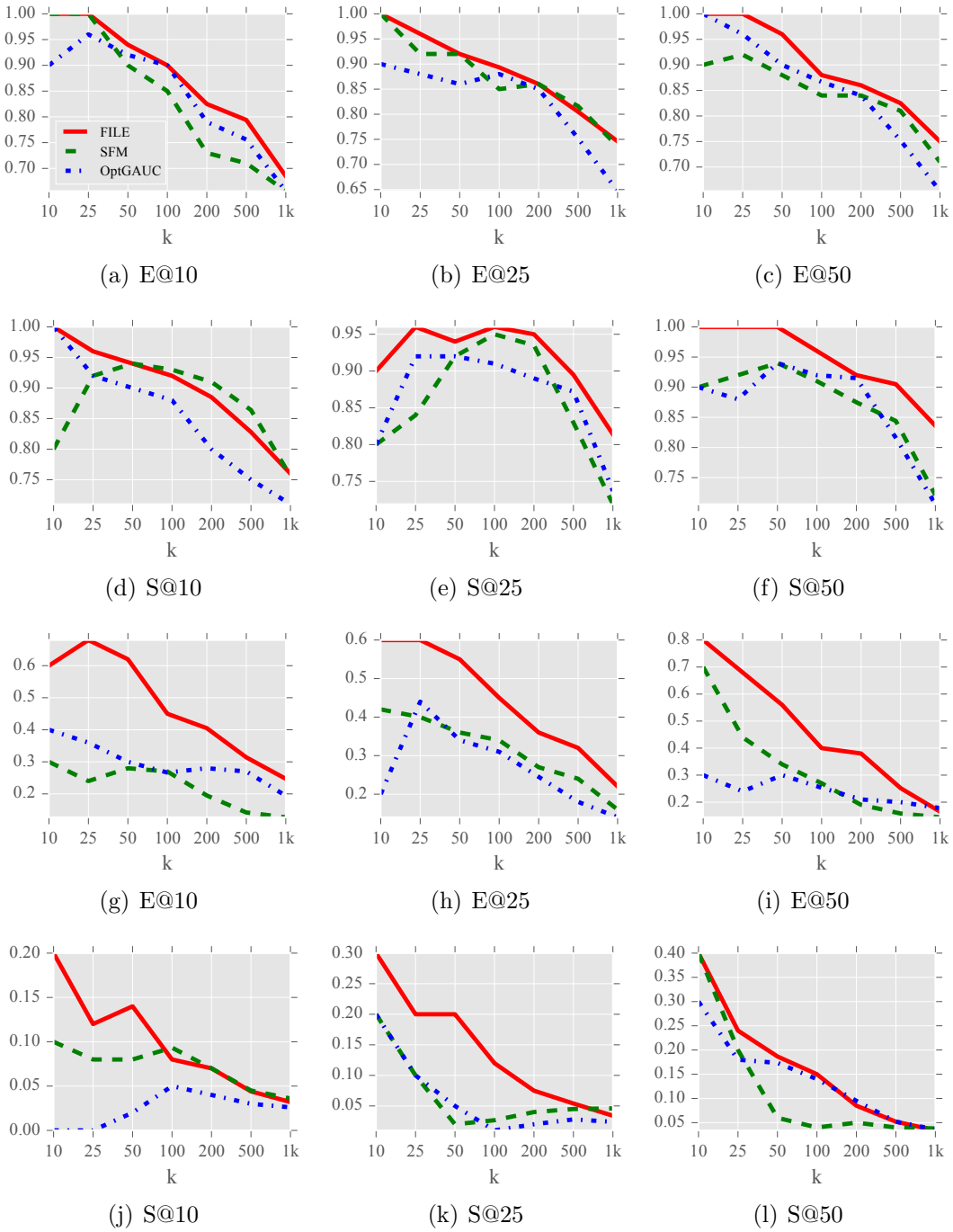
FIGURE 4.4: Performance fluctuations across datasets with different parameter combinations.

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 4.3. 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 4.4. 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.

4.3.2 Comparative Experiments

4.3.2.1 Overall Performance.

Table 4.4 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



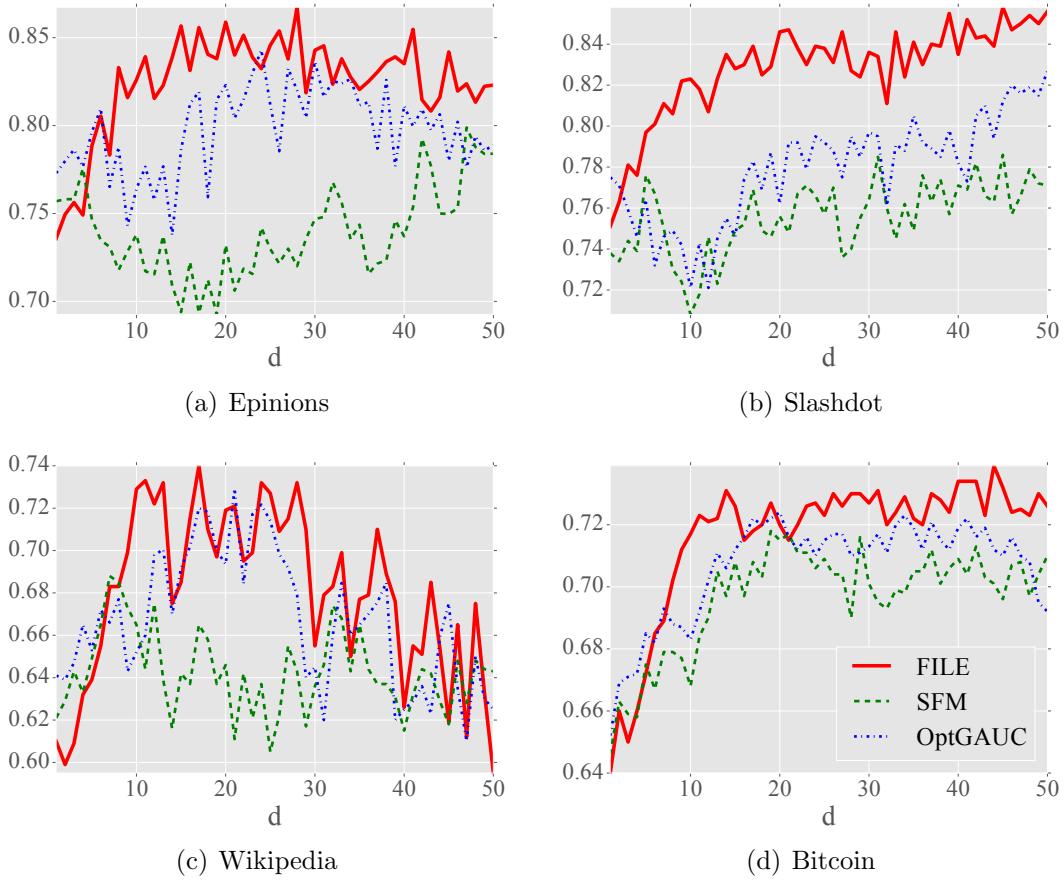
(a) E@10 (b) E@25 (c) E@50

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



(a) Epinions (b) Slashdot

(c) Wikipedia (d) Bitcoin

FIGURE 4.6: The impact of degree d.

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 [11]. 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 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).

4.3.2.2 Top-k Ranking Performance.

We investigate the ranking performance on top k. Both precision of positive (i.e., PPrec) and negative (i.e., NPrec) at top k (k=$\{10, 25, 50, 100, 200, 500, 1000\}$) are examined. We show the experimental results in six datasets for each metric in Figure 4.5, 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 practical and effective in real-world applications. It is worth noting that, FILE exhibits greater improvement over other approaches in terms 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.

4.3.2.3 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 4.6, 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.

4.3.3 Application: fraudulent user detection

Fraud detection is a crucial application in the security domain. In social networks, fraudulent users give fake content or ratings and eventually get distrust from others. Fraud detection task aims to distinguish fraudulent users from benign ones, which takes up the majority. For a user in signed networks, the components of its received links indicate its trustworthiness: a user will be trustworthy if it receives more positive links than negative ones. Therefore, in this experiment, we define fraudulent users are the ones who receive more negative links than positive links, while a user is defined as benign if it receives at least 10 more positive links than negative links. We choose AUC as the metric because the size of fraudulent users and benign users are always imbalanced. In this application, we aim to distinguish fraudulent users from benign users, which is a classification problem over two classes. Therefore, we adopt AUC instead of GAUC since GAUC is used to distinguish three classes. For a fair comparison, we set r=10 and d=50 for all baselines.

We adopt random forest as the classifier, and the comparison results are listed in table 4.5. We can see that existing models cannot perform well consistently while our model performs better than baselines. The good performance indicates that latent features derived from FILE model well represent users' intrinsic character and can be applied to fraudulent detection for online security.

TABLE 4.5: Fraudulent user detection performance

Models	Epinions	Slashdot
LRM	0.863	0.746
BPRMF	0.812	0.785
OptGAUC	0.872	0.788
FILE	**0.881**	**0.803**

4.4 Summary

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 chapter, 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.

Chapter 5

SSRW: Supervised User Ranking in Signed Networks[1]

In Chapter 4, the research question we aim to answer is: give two user pairs in signed networks, which pair is more likely to become friends or enemies? Specifically, the output of the models is the ranking list of user pairs, by the order of positive connection, no-relation and negative connection. Therefore, these works focus on generating a global ranking list for the whole network, which could easily lead to a relatively unfair scenario where some users might have a large number of potential links in the ranking list while most users have very few or even no links. In this case, they cannot be easily adapted for many real-world applications such as social recommendation or social-aware product recommendation. In contrast, personalized user ranking, which generates a ranking list for each individual, is more practical and realistic [29]. Besides, the ranking list for a user provided by existing random walk methods is fixed given a certain network snapshot (i.e. the network structure). They inappropriately assume all the links have the same weights (i.e. social strengths, a.k.a. link strengths). In other words, they cannot

[1]The work in this chapter has been published as [28] in AAAI 2019.

learn each individual's own opinions towards her neighbors, such as what kind of user link (i.e. neighbors) is more important.

To fill the research gap, we propose Signed Supervised Random Walk (SSRW), through which we learn social strengths that capture a user's different preferences towards different neighbors, and thus to better facilitate the task of personalized user ranking. More specifically, instead of considering the random walk in a given network snapshot (i.e. training data), we split the training data into two parts in terms of the timestamp (denoted as A and B), and learn social strengths (i.e. transition probabilities) so that random walk more likely visits those newly positively connected nodes (i.e. in B compared to A) whereas more reluctantly visits the newly negatively connected nodes. We conduct experiments on four real-world datasets and the results show that SSRW's performance has an improvement of 6.05% compared to state-of-the-art approaches. To improve SSRW's efficiency but simultaneously maintain its effectiveness, we also design a fast ranking method (F-SSRW) based on the local structure among each seed node and a certain set of candidates of the seed node. It has been demonstrated that F-SSRW can maintain the performance in contrast with the original SSRW when the ranking candidates of a user satisfy the requirement of having substantial common neighbors with the user.

5.1 Problem Formulation and Transformation

We first define the user ranking problem as: given a seed node i in a signed social network $S \in \mathbb{R}^{n \times n}$ (n is the number of users) with $S_{ij} \in \{1, 0, -1\}$, we aim to rank all the users $m \in \{m | S_{im} = 0\}$ in the present, by the probability of transforming (i, m) to a positive link, maintaining no-relation, or transforming to a negative

link in the future. We aim to answer the question: "*Of user pairs (i, m_1) and (i, m_2), which pair is more likely to become friends (or enemies)?*"

As aforementioned, social strengths have been ignored by existing approaches in the literature. In fact, the intuition that a user's preferences towards other users (even towards the set of already formed friends) are different, has been widely explored and leveraged in the unsigned networks [77, 78]. We thus adopt the idea and consider that social strengths can also impact link formation in signed networks. Therefore, we transform the user ranking problem into a supervised learning problem, by which we learn social strengths to better facilitate user ranking task.

Formally, for any link (i, j), we learn its link strength $f_w(x_{ij})$, in which $f(\cdot)$ is a differentiable function parameterized by w and x_{ij} is the observable feature vector of the link. By doing this, we obtain a weighted network with different edge strength $f_w(x_{ij})$. We use r_{im} to represent m's ranking score, which is the probability obtained from random walk based on the weighted network. Thus, the problem is reduced to: "*given a seed node i and any nodes $m, n \in C_i$, we aim to find the optimal w, which satisfies: if there is a new positive link generated from i to m in future, the ranking score should follow $r_{im} \geq r_{in}$. Similarly, $r_{im} \leq r_{in}$ if a negative link is formed between i and m.*"

The main notations are summarized in Table 5.1. For the seed user i, we aim to optimize the following function:

$$\begin{aligned} \underset{w}{\text{Minimize}} \quad & F(w) = ||w||^2 + \frac{1}{\theta} \sum e_{mn} \\ \text{s.t.} \quad & r_{in} - r_{im} + e_{mn} \geq 1, \ \forall m, n \in C_i \end{aligned} \tag{5.1}$$

TABLE 5.1: Notations.

S	adjacency matrix, $S \in \mathbb{R}^{n \times n}$, $S_{ij} \in \{1, 0, -1\}$
i	seed node
(i, j)	link $i \to j$
C_i	node i's candidate set, $C = P \cup U \cup N$
m, n	node $m, n \in C_i$
x_{ij}	feature vector of (i, j)
$f_w(x_{ij})$	strength function of (i, j) parameterized by w
P_i	$P_i = \{m \vert S_{im}^t = 0 \ \& \ S_{im}^{t+1} = 1, m \in C_i\}$
N_i	$N_i = \{m \vert S_{im}^t = 0 \ \& \ S_{im}^{t+1} = -1, m \in C_i\}$
U_i	$U_i = \{m \vert S_{im}^t = 0 \ \& \ S_{im}^{t+1} = 0, m \in C_i\}$
δ_i	node i's bias on distrust
r_{im}^+	seed node's positive ranking score towards m
r_{im}^-	seed node's negative ranking score towards m
r_{im}	the final ranking score
q	unit vector with $q_i = 1$

where $\theta = |N_i| \cdot |P_i| + |U_i| \cdot |P_i| + |N_i| \cdot |U_i|$, in which $|P_i|$, $|U_i|$ and $|N_i|$ are the number of nodes in the corresponding set respectively, and $\sum e_{mn}$ is equivalent to:

$$\alpha \sum_{m \in N_i, n \in P_i} e_{mn}^1 + \beta \sum_{m \in U_i, n \in P_i} e_{mn}^2 + \gamma \sum_{m \in N_i, n \in U_i} e_{mn}^3 \tag{5.2}$$

The corresponding weights α, β, γ are user-specified and application-dependent, denoting the penalties of different types of errors, where e_{mn}^1 is type 1 error that $m \in N_i, n \in P_i$, e_{mn}^2 is type 2 error that $m \in U_i, n \in P_i$, and e_{mn}^3 is type 3 error that $m \in N_i, n \in U_i$. The objective of the optimization equation 5.1 is to find the optimal parameter set w and can be proceeded once the ranking score r and $\frac{\partial r}{\partial w}$ are obtained. Therefore, our main research question is reduced to "how to design the function r and then calculate its derivation accordingly". In this chapter, we extend the supervised random walk technique [79] to signed networks, i.e. *signed supervised random walk*, for obtaining the ranking score r.

5.2 SSRW: Signed Supervised Random Walk

In SSRW, we first follow the idea of the sign surfer [29] to make random walk workable in signed networks. A surfer begins with a positive sign (since it will always trust itself) and then visits other nodes, and the sign flips if it meets a negative link, otherwise the sign remains unchanged.

The intuition of the sign flips is adopted from balance theory [64], which can be explained as "my friend's friend is my friend" or "my enemy's friend is my enemy", and considered solidly effective in signed networks [7]. As a personalized ranking approach, the surfer will restart with a probability c, and the sign will be reset to positive. When the surfer visits a certain node, the sign can be either positive or negative since it can reach the node via different routes. Therefore, for the seed node i, each node m in C_i will eventually get two ranking scores: a positive one (i.e. r^+_{im}) and a negative one (i.e. r^-_{im}), from which we can obtain the final ranking score r as:

$$r_{im} = r^+_{im} - \delta_i r^-_{im} \tag{5.3}$$

where δ is user i's bias on distrust (i.e. negative relationship), as some users will be more likely to distrust others, while some might be more reluctantly to distrust others. Next, we investigate the connection between link strength $f_w(\cdot)$ and r. Let a_{ij} be the normalized link strength of (i, j), which equals to 0 if there is no direct link from node i to j:

$$a_{ij} = \frac{f_w(x_{ij})}{\sum_z f_w(x_{iz})} \ , \exists(i,j) \tag{5.4}$$

In this case, the matrix of transition probability Q is:

$$Q_{ij} = \begin{cases} a_{ij} & \exists(i,j) \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

and we split Q to two matrices Q^+ and Q^- according to the link sign between users. Specifically, $Q_{ij} \in Q^+$ if $S_{ij} = 1$, and $Q_{ij} \in Q^-$ if $S_{ij} = -1$. We thus have:

$$Q = Q^+ + Q^- \tag{5.6}$$

Based on the setting and sign surfer, seed node i's ranking score matrices r_i^+ and r_i^- towards other users are recursively entangled and derived as follows:

$$\begin{aligned} r_i^+ &= (1-c)(Q^+ r_i^+ + Q^- r_i^-) + cq \\ r_i^- &= (1-c)(Q^+ r_i^- + Q^- r_i^+) \end{aligned} \tag{5.7}$$

where q is the unit vector with $q_i = 1$. Thus, for a node $m \in C_i$, its ranking score can be written as:

$$\begin{aligned} r_{im}^+ &= (1-c)\textstyle\sum_j (r_{ij}^+ Q_{jm}^+ + r_{ij}^- Q_{jm}^-) \\ r_{im}^- &= (1-c)\textstyle\sum_j (r_{ij}^- Q_{jm}^+ + r_{ij}^+ Q_{jm}^-) \end{aligned} \tag{5.8}$$

We take the corresponding derivations to compute $\frac{\partial r_{im}}{\partial w}$:

$$\begin{aligned} \frac{\partial r_{im}^+}{\partial w} &= (1-c)\textstyle\sum_j Q_{jm}^+ \frac{\partial r_{ij}^+}{\partial w} + r_{ij}^+ \frac{\partial Q_{jm}^+}{\partial w} + Q_{jm}^- \frac{\partial r_{ij}^-}{\partial w} + r_{ij}^- \frac{\partial Q_{jm}^-}{\partial w} \\ \frac{\partial r_{im}^-}{\partial w} &= (1-c)\textstyle\sum_j Q_{jm}^+ \frac{\partial r_{ij}^-}{\partial w} + r_{ij}^+ \frac{\partial Q_{jm}^-}{\partial w} + Q_{jm}^- \frac{\partial r_{ij}^+}{\partial w} + r_{ij}^- \frac{\partial Q_{jm}^+}{\partial w} \end{aligned} \tag{5.9}$$

where $\frac{\partial Q_{jm}}{\partial w}$ can be derived as:

$$\frac{\partial Q_{jm}}{\partial w} = \frac{\frac{\partial f_w(x_{jm})}{\partial w}(\sum_z f_w(x_{jz})) - f_w(x_{jm})(\sum_z \frac{\partial f_w(x_{jz})}{\partial w})}{(\sum_z f_w(x_{jz}))^2} \tag{5.10}$$

Similarly, r^+_{im}, r^-_{im}, $\frac{\partial r^+_{im}}{\partial w}$, and $\frac{\partial r^-_{im}}{\partial w}$ are recursively entangled, and can be computed iteratively as in Algorithm 1.

Algorithm 2 Computation of r^+_{im}, r^-_{im}, $\frac{\partial r^+_{im}}{\partial w}$, and $\frac{\partial r^-_{im}}{\partial w}$

Initialize: $r^{+(0)}_{im}$, $r^{-(0)}_{im}$, $\frac{\partial r^{+(0)}_{im}}{\partial w} = 0$, $\frac{\partial r^{-(0)}_{im}}{\partial w} = 0$,
$t = 1$
repeat
  Calculate $r^{+(t)}_{im}$, $r^{-(t)}_{im}$ based on Equation 5.8;
  $t = t + 1$;
until Converge
$r^+_{im} = r^{+(t-1)}_{im}$, $r^-_{im} = r^{-(t-1)}_{im}$
$t = 1$
repeat
  Calculate $\frac{\partial r^{+(t)}_{im}}{\partial w}$, $\frac{\partial r^{-(t)}_{im}}{\partial w}$ based on Equation 5.9;
  $t = t + 1$;
until Converge
$\frac{\partial r^+_{im}}{\partial w} = \frac{\partial r^{+(t-1)}_{im}}{\partial w}$, $\frac{\partial r^-_{im}}{\partial w} = \frac{\partial r^{-(t-1)}_{im}}{\partial w}$
Output: r^+_{im}, r^-_{im}, $\frac{\partial r^+_{im}}{\partial w}$, $\frac{\partial r^-_{im}}{\partial w}$

Based on the computed r^+_{im}, r^-_{im}, $\frac{\partial r^+_{im}}{\partial w}$, and $\frac{\partial r^-_{im}}{\partial w}$, we then apply stochastic gradient descent (SGD) [80] to find a local minimum for Equation 5.1.

The advantage of SSRW is to well capture the global structure of a network, and thus to obtain the global optimum, but it is also computationally expensive. Through SSRW, every single user $m \in \{m|S_{im} = 0\}$ will eventually get a ranking score with respect to the seed node i. However, it is intuitive that most users will not connect to the seed node eventually. In other words, we do not need to consider all the nodes in the network, since the link formation probabilities between the seed node and them are negligible. Inspired by this, we thus propose a fast ranking method, denoted as F-SSRW and detailed in the next section.

5.3 F-SSRW

As aforementioned, given a seed node, we aim to speed up the ranking procedure by only considering a certain set of candidates who have a much larger probability to form a link with the seed node. Intuitively, we select candidates by two criteria: (1) hop distance between the seed node and the candidates; (2) the number of their mutual neighbors.

In this section, we first empirically investigate these two factors which may influence link formation. We mainly conduct the analysis in Epinions dataset[2], which contains the timestamp of every link formation over 30 months. We also perform analysis in other three real-world signed networks: Slashdot[3], Wikipedia RFA[4] and Bitcoins[5].

5.3.1 Hop distance

We check the hop distance between two users when they are linked. As can be seen in Table 5.2, two-hop distance is dominant with which a much larger number of links are configured compared to all other ones. Besides, as shown in Figure 5.1(a), the majority of connected links have direct common neighbors (i.e. the corresponding two users are within two-hop distance). In other words, users are more likely to get linked if they are within two-hop distance. This intuition has also been well validated in unsigned networks [81].

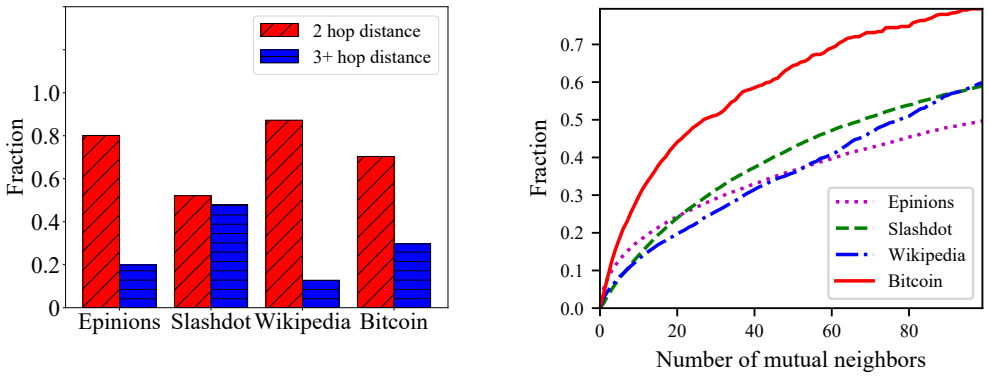
[2]http://www.trustlet.org/epinions.html
[3]snap.stanford.edu/data/soc-Slashdot0902.html
[4]snap.stanford.edu/data/wiki-RfA.html
[5]cs.umd.edu/~srijan/wsn

TABLE 5.2: The hop distance between users when they form new links in Epinions dataset

Hop distance	Link counts	Ratio
2	187,990	71.65%
3	32,014	12.20%
4+	42,371	16.15%

5.3.2 The number of mutual neighbors

We further check the relation between link formation and the number of mutual neighbors. Figure 5.1(b) depicts the cumulative distribution of the number of mutual neighbors between linked users. Specifically, more than 90% of links have at least 3 mutual neighbors, and more than 82% of them have at least 5 mutual neighbors. Therefore, in signed networks, users with more common neighbors are more likely to form links (either positive or negative).



(a) (b)

FIGURE 5.1: (a) Distribution of hop distance; (b) Distribution of the number of mutual neighbors.

The results of data analysis further inspire our design of the fast ranking model which mainly focuses on local structure among the seed node and certain set of candidates. Specifically, for a seed node i, we prune the graph to include only $\{i, \{j\}, m\}$, in which j is a common neighbor of i and m, and $j \in \{j|\exists(i,j)\ \&\ (j,m)\}$.

Besides, we only select candidate nodes from C_i which have at least T mutual neighbors with i. In this case, we keep the candidates which are more likely to obtain a higher r_{im}^{+} or r_{im}^{-}. In the pruned graph, based on random walk, a candidate node m's ranking score r_{im} can be estimated as:

$$r_{im} \propto \sum_{j} r_{ij} a_{jm} \tag{5.11}$$

Because the pruned graph only contains i's two-hop neighbors, we can further approximate r_{im} as

$$r_{im} \propto \sum_{j} r_{ij} a_{jm} \propto \sum_{j} a_{ij} a_{jm} \tag{5.12}$$

where we strategically ignore the routes with more than 2 hops. We argue that the gap between the approximation and global optimum (obtained from random walk) is actually small since the contribution of 3-hop routes is limited in term of hitting probabilities compared to 2-hop routes. The intuition has been shown in data analysis meanwhile being well validated in the literature [81].

Similarly, grounded on balance theory, we will also obtain two ranking scores, r_{im}^{+} and r_{im}^{-} in the signed network:

$$\begin{cases} r_{im}^{+} \propto \sum a_{ij} a_{jm} & \text{if } S_{ij} S_{jm} = 1 \\ r_{im}^{-} \propto \sum a_{ij} a_{jm} & \text{if } S_{ij} S_{jm} = -1 \end{cases} \tag{5.13}$$

Finally, we obtain r_{im} following Equation 5.3, denoted as:

$$r_{im} = \sum_{j} (r_{im}^{j+} - \delta r_{im}^{j-}) \tag{5.14}$$

where r_{im}^{j+} and r_{im}^{j-} denote the respective ranking score contributed by the route $i \to j \to m$, and following Equation 5.4:

$$r_{im}^{j} = \frac{f_w(x_{ij}) \cdot f_w(x_{jm})}{\sum_k f_w(x_{ik}) \sum_k f_w(x_{jk})} \tag{5.15}$$

By taking derivation on equation 5.14, we thus obtain:

$$\frac{\partial r_{im}}{\partial w} = \sum_j (\frac{\partial r_{im}^{j+}}{\partial w} - \delta \frac{\partial r_{im}^{j-}}{\partial w}) \tag{5.16}$$

Based on the computed r_{im} and $\frac{\partial r_{im}}{\partial w}$, we apply stochastic gradient descent to find a local minimum for Equation 5.1.

5.3.3 Discussion

F-SSRW is a simplified version of SSRW, which works on the pruned graph and only focuses on selected candidates. Therefore, in F-SSRW, we can only obtain approximate ranking scores for these candidates. In contrast, SSRW works on the global graph and every single user $m \in \{m|S_{im} = 0\}$ will eventually get a ranking score. For both of them, the time complexity of the optimization method is $O(|N_i| \cdot |P_i| + |U_i| \cdot |P_i| + |N_i| \cdot |U_i|)$. For each iteration to get the derivation, SSRW takes $O(|E|+|V|)$, in which $|E|$ is the number of links and $|V|$ is the number of nodes in the graph, whereas in each iteration F-SSRW takes $O(|C_i|)$. As the candidate set and the graph in F-SSRW are much smaller than those in SSRW, accordingly the efficiency of the F-SSRW is much more largely improved than that in SSRW.

TABLE 5.3: Dataset statistics.

		Epinions	Slashdot	Wikipedia	Bitcoin
	Users	131,828	82,140	9,654	3,783
	Positive links	717,667	425,072	87,766	22,650
	Negative links	123,705	124,130	16,788	1,536
$d \geq 3$	Candidate users	800.4	137.4	243.9	201.1
(Per seed)	Linked users	54.4	9.8	34.1	29.9
$d \geq 5$	Candidate users	382.3	41.6	106.5	85.6
	Linked users	44	5.7	23.8	20.6
$d \geq 10$	Candidate users	168.4	14.9	22.9	26.7
	Linked users	31.9	3.5	9	11.4

5.4 Experiments

We conduct experiments on four real-world datasets and compare the proposed approaches with the state-of-the-art methods.

5.4.1 Experimental Settings

5.4.1.1 Data.

We employ the four datasets (i.e. Epinions, Slashdot, Wikipedia RFA and Bitcoin), which are the only public available datasets with signed structure. In this study, we focus on the users (i.e. seed nodes) who are active in the social networks, where the activeness is measured by user's degree. Specifically, to conform with the previous studies [79], the selected users' degree is larger than 20. We randomly select 200 of them as seed nodes. Besides, to make a more comprehensive evaluation, we adopt different criteria for candidate node selection. Based on the data analysis, we only select two-hop neighbors as candidates, meanwhile further filter them by the number of mutual neighbors with the seed node. Specifically, we use $d \geq 3$, $d \geq 5$ and $d \geq 10$. The major statistics of the datasets are listed in Table 5.3.

TABLE 5.4: Performance of different methods. The best performance is highlighted in bold, and the second-best one (except SSRWs) is marked by *. 'Improvement' indicates the improvement of SSRW over the model having the highest performance among existing models.

Datasets	**SPNR**	**TNS**	**SFM**	**RWR**	**G-SSRW**	**F-SSRW**	**SSRW**	**Improvement**
Epinions@3	0.619	0.462	0.609	0.628*	0.644	0.639	**0.678**	7.96 %
Epinions@5	0.621	0.475	0.633	0.671*	0.660	0.676	**0.702**	4.62 %
Epinions@10	0.598	0.499	0.677*	0.647	0.673	0.729	**0.743**	9.75 %
Slashdot@3	0.619	0.567	0.541	0.633*	0.601	0.626	**0.645**	1.90 %
Slashdot@5	0.605	0.563	0.537	0.627*	0.612	0.634	**0.659**	5.10 %
Slashdot@10	0.554	0.628	0.569	0.645*	0.658	**0.718**	0.715	11.32 %
Wikipedia@3	0.549	0.545	0.487	0.568*	0.583	0.587	**0.633**	11.44 %
Wikipedia@5	0.544	0.582	0.574	0.579*	0.612	0.619	**0.638**	10.38 %
Wikipedia@10	0.558	0.598	0.679*	0.596	0.647	0.658	**0.681**	0.29 %
Bitcoin@3	0.589	0.472	0.554	**0.613**	0.574	0.588	0.601*	−0.33 %
Bitcoin@5	0.596*	0.490	0.585	0.583	0.599	0.601	**0.614**	6.67 %
Bitcoin@10	0.573	0.557	0.640*	0.615	0.621	**0.663**	0.657	3.59 %

5.4.1.2 Evaluation Metrics.

We use 2-fold cross-validation for training and testing, and utilize GAUC (Generalized AUC) [11] to measure the ranking performance.

Another metric is precision@top k, by which we evaluate the performance of the link recommendation. Specifically, we use PPrec@k (NPrec@k) to denote the ratio of positive (or negative) links in the top (or bottom) k prediction.

5.4.1.3 Benchmarking approaches.

We make comparisons with state-of-the-art approaches, including similarity-based models: Similarity with Positive and Negative Relations (SPNR) [23], friend Transitive Node Similarity (TNS) [22], Social Feature Model (SFM) [26]; and random walk based model: Signed Random Walk with Restart (RWR) [29]. We also compare our personalized model with the global version of our model (G-SSRW), which is based on SSRW but strives to minimize the sum of losses over all seed nodes in the network.

5.4.1.4 Parameter settings.

In this experiment, we consider 5 features for the vector x to describe a user pair. Two of the features are two users' degrees respectively, which imply their activeness; and the rest three are the number of their common friends, enemies, and frenemies (one's friend but the other one's enemy) respectively, which describe the social patterns within their joint relationship. we utilize the linear model to represent the link strength, i.e., $f_w(x) = w^T x$, where w can be seen as the weight vector of the features, and denote importance degrees of the corresponding features.

For the benchmark approaches, we set the parameters recommended in the literature. In SSRW, there are five hyper-parameters: δ, α, β, γ and the restart probability c. The first four are application-dependent, and in view of simplicity and fair comparisons, we make them equal to 1 respectively. Besides, we set $c = 0.2$ for SSRW in the comparative experiments considering favourable performance of our model in this setting.

5.4.2 Experimental Results

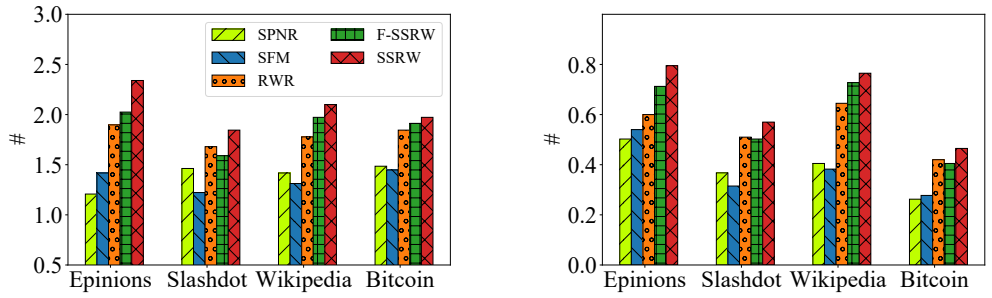
Here, we show the comparison results under different scenarios and the impact of different parameters on the approaches.

5.4.2.1 Overall Performance

Table 5.4 depicts the experimental results under different scenarios in terms of GAUC. Overall, SSRW achieves the best performance when compared with other approaches across all the datasets, and the improvement is 6.05% on average. The results of t-test demonstrate that the improvement of our approach is statistically significant (p-value$<$ 0.01).

Particularly, among all these approaches, similarity-based models (SPNR and TNS) perform the worst under almost all scenarios, indicating that traditional similarity metrics cannot be easily extended into signed networks. The global ranking approaches, including SPNR, TNS and SFM, perform worse than the personalized approaches (e.g. RWR). However, SSRW and F-SSRW perform much better than RWR, validating the effectiveness of the supervised approach and personalized link strengths. Besides, SSRW performs much better than G-SSRW, implying the reasonability of the personalized user ranking compared to the global user ranking.

With regard to the SSRW and F-SSRW, we can see that SSRW is better than F-SSRW, but F-SSRW performs almost better than all the rest approaches except SSRW. In addition, as the increase of d, the performance gap between F-SSRW and SSRW becomes smaller, further demonstrating the effectiveness of our heuristic intuitions.

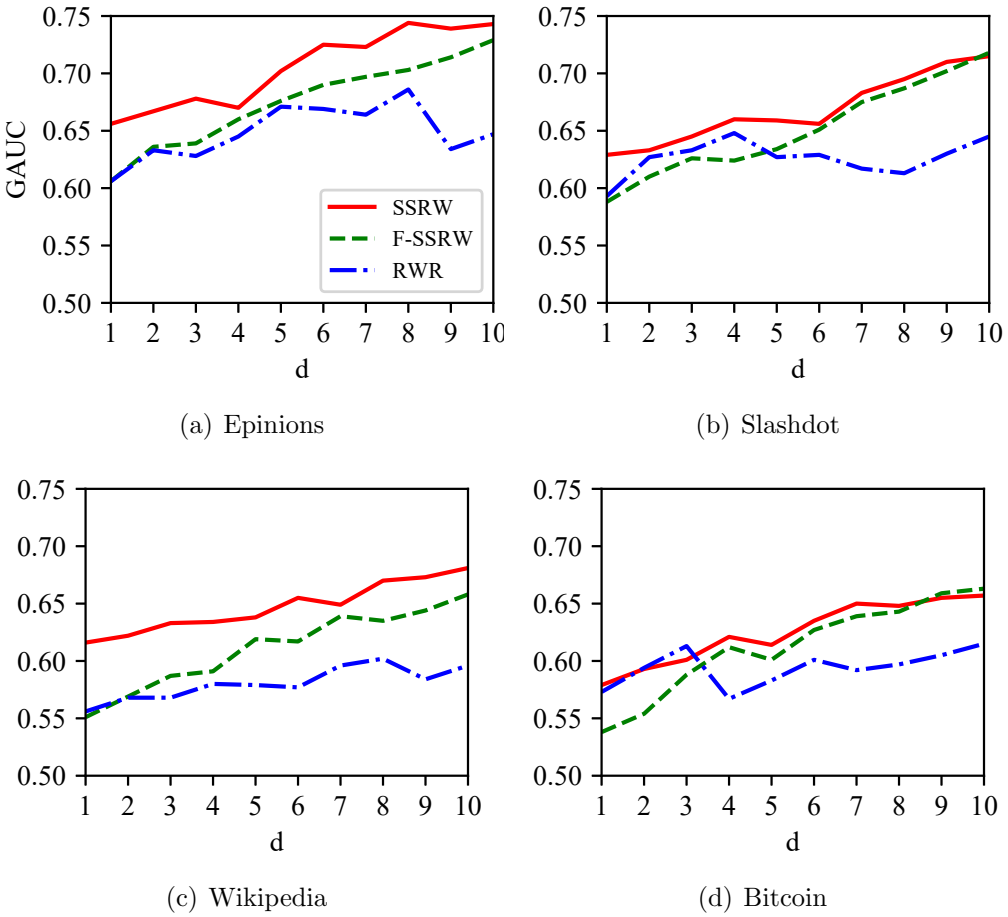


(a) PPrec@10 (b) NPrec@10

FIGURE 5.2: Comparative performance in terms of ranking top 10 positive links and negative links.

5.4.2.2 Impact of candidate selection by d

To demonstrate the robustness of the proposed approaches, we check the performance of different approaches in terms of GAUC as the change of d in the range of $[1, 10]$. We compare SSRW and F-SSRW with RWR as it performs the best among all the benchmarks. As shown in Figure 5.3, we can find that SSRW consistently performs better than RWR and F-SSRW. As d increases, the performance gap between F-SSRW and SSRW becomes smaller, validating the soundness of our argument that a greater d can assure a better approximation of F-SSRW compared to SSRW. In other words, considering the efficiency, we can adopt the F-SSRW model in those applications where the candidate nodes have substantial common neighbors with the seed node.



(a) Epinions (b) Slashdot

(c) Wikipedia (d) Bitcoin

FIGURE 5.3: The comparative performance with the change of d.

5.4.2.3 Runtime comparison

We then further empirically check the actual runtime of the experiments conducted on a four CPU 3.7GHz machine with 16GB memory. Figure 5.4 shows the runtime comparison between SSRW and F-SSRW on the Epinions dataset, and we can see that F-SSRW is significantly efficient than SSRW.

5.4.2.4 Precision@Top-k

We investigate the ranking performance of different approaches in terms of PPrec@k and NPrec@k. Figure 5.2 shows the comparison results on top 10 precision when d equals to 3. We can see that SSRW consistently achieves the best results across

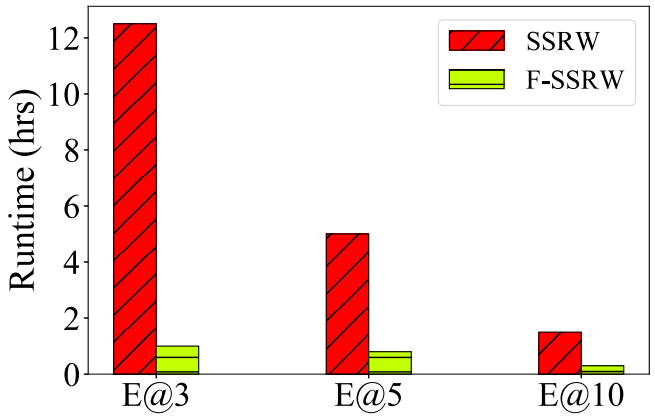


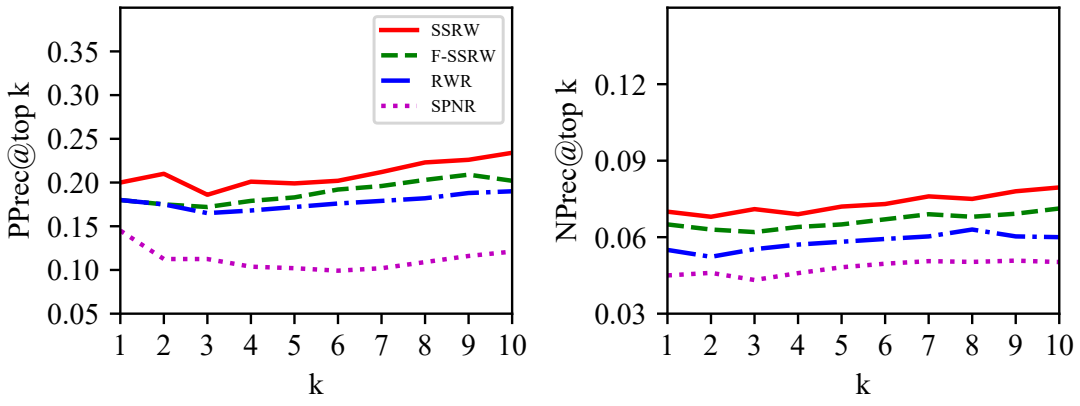
FIGURE 5.4: Runtime comparison.

all datasets. Random-walk based approaches (i.e. SSRW, F-SSRW and RWR) obtain better performance than other benchmarks, implying that simply taking local attributes into consideration for similarity-based metrics cannot assure satisfying performance in personalized user ranking task. The better performance of SSRW and F-SSRW compared with others also indicates the reasonability of taking personalized social strengths into account.

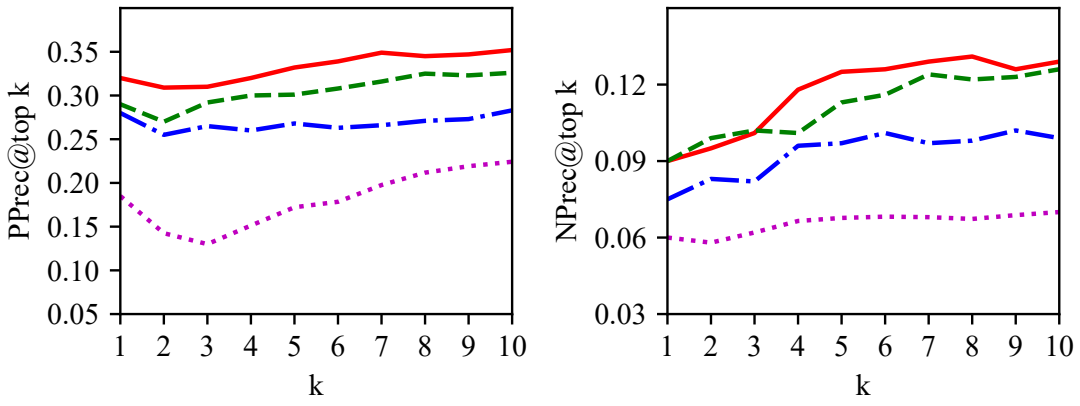
We also examine the performance of top k by varying k in the range $[1, 10]$, along with different d. We show the experimental results in Epinions in Figure 5.5, which demonstrate the consistent superior of SSRW over benchmarks. Besides, the performance of F-SSRW becomes better as d increases. Overall, the results imply the effectiveness of our model on top@k ranking, where positive top @k can be used for link recommendation, whereas the negative top k can be used in security-related domains.

5.4.2.5 Impact of the parameter c

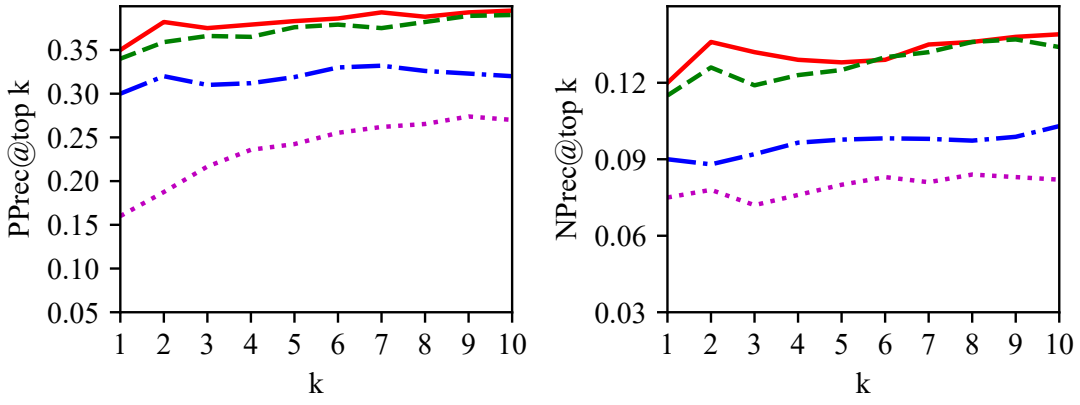
The restart probability c is an important parameter for random walk. A smaller c will allow the model 'walk' far away from the seed node while a larger c will force the model to walk within the local structure. We thus check the impact of c



(a) $d = 3$



(b) $d = 5$



(c) $d = 10$

FIGURE 5.5: PPrec@top k (left) and NPrec@top k (right) in the Epinions dataset with different d.

on SSRW in terms of GAUC by varying d in the range of $[0.1, 0.9]$. As shown in Figure 5.6, c indeed affects SSRW's performance, and we can obtain a relatively better performance when $c \in [0.2, 0.4]$. When $c \geq 0.4$, SSRW performs slightly worse with the increase of d. However, the performance variance is insignificant, indicating that SSRW is relatively insensitive and robust in terms of the restart probability c.

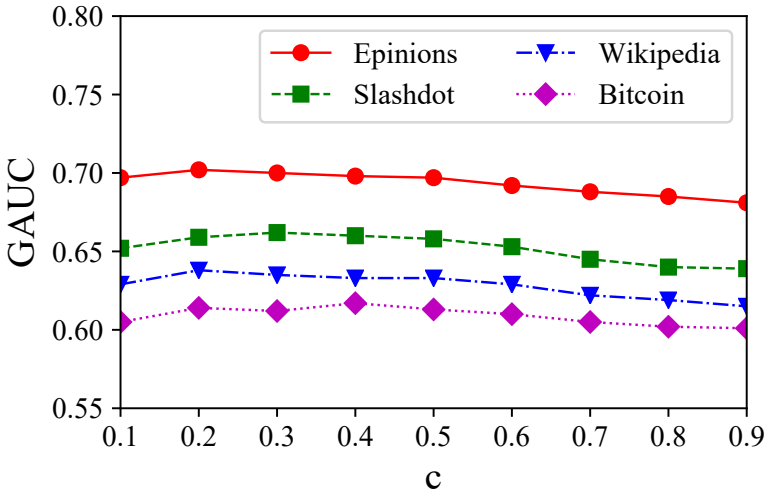


FIGURE 5.6: Impact of the parameter c on SSRW.

5.5 Summary

User ranking is a fundamental and key research problem in signed networks, which has wide applications in real-world scenarios such as recommendation systems and security-related platforms. In this chapter, we propose the SSRW model which learns social strengths to optimize the user ranking list for each individual user. Specifically, we apply supervised random walk in signed scenarios and learn link strengths to guide more effective random walk. Based on the heuristics from data analysis, we further design a simplified and efficient ranking method (F-SSRW), which only focuses on certain candidate nodes and runs the learning algorithm within the local graph of the seed node. A comprehensive evaluation demonstrates

the superiority of the proposed models over state-of-the-art approaches, and the robustness in terms of parameters and experimental settings. There are mainly two limitations in this study that could be addressed in future research. First, the computation cost of SSRW model is expensive. The model needs to be further improved in order to apply on the platforms with large groups of users. Second, in this work, we use the linear function to represent the social strength with 5 explicit features as variables. More features can be adopted to enhance the prediction performance.

Chapter 6

Conclusions and Future Works

In this thesis, we focus on link prediction problems in signed networks, and we have proposed a feature-based model, a latent model, and a supervised random walk based model. In this chapter, we summarize the contributions of this thesis and discuss several promising directions for future works.

6.1 Conclusions

Link prediction in signed networks is fundamental research and its goal is to predict which unconnected user pairs will be linked in the future and also predict the link sign. Existing approaches treat it as a classification problem, to solely predict the link sign, with the assumption of the link existence. Therefore, in this thesis, we aim to have a deep understanding on the link formation mechanism in signed networks and thus develop prediction models.

To reveal the underlying mechanism regarding link formation in signed networks, we propose a structured feature framework on the basis of a thorough feature analysis. We propose a structured feature framework on the basis of a thorough

feature analysis to reveal the underlying mechanism regarding link formation in signed networks. We not only adopt existing features in previous studies [7, 30] on both unsigned and signed network scenarios, but also derive new features based on social theories and observations. The feature framework, grounded on both well-known theories and sound observations, can serve as a guidance for research on the new problem.

Chapter 4 makes a deep investigation on the no-relation status. We empirically show that two types of no-relation status widely exist in real-world social networks. We also find that the link prediction problem in signed networks becomes rather difficult mainly due to the diversity of no-relation, as it is easy to mispredict no-relation having many common neighbors as a linked status. Therefore, 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 then propose the FILE 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.

Finally, different from the FILE framework which aims to optimize the global ranking performance, Chapter 5 focus on the personalized user ranking problem. We apply the supervised random walk in signed network scenario, and make users more likely visit their "potential friends" and less likely visit their "potential enemies". In this work, we propose a supervised ranking model which learns social strengths to optimize the user ranking in signed networks. A comprehensive evaluation demonstrates the superiority of the proposed models over state-of-the-art approaches, and the robustness in terms of parameters and experimental settings.

6.2 Future Works

Our works mainly focus on understanding link formation in signed social networks. Based on these studies, we propose some potential future research directions.

6.2.1 Extending the proposed models

First, our feature-based model can be enhanced by considering more social theories like Emotional theory [82, 83]. Besides, we will investigate more real-world datasets to further evaluate the significance of the new problem and the effectiveness of our approach. Based on specific scenarios and social theories, we can explore more features for the feature-based model.

Second, our FILE framework works well even only two explicit features are adopted. To improve the FILE framework, we can explore more explicit features to enhance its prediction performance, and further test the effectiveness of it using field experiments. Besides, we have applied the model in the security domain like fraudulent user detection, in the future, we may use it for social recommendation task.

Third, to improve the SSRW model, we will try more complex functions to represent the ranking score function by simultaneously incorporating more explicit features. Besides, we strive to apply SSRW in real-world scenarios such as social recommendation to further validate their effectiveness.

6.2.2 Applying deep learning techniques

Recently years, deep learning techniques have been applied in almost every corner of the machine learning domain, including target detection [84] and image classification [85]. Current research [86] has shown its effectiveness in feature learning.

Therefore, we may adopt deep learning to learn the topological features in signed networks, thus to enhance link prediction performance. Specifically, we may aim to learn the features from the social connections between nodes by neural network. It's a promising way since neural network has a super expressing power compared to the existing approaches.

6.2.3 Understanding the "link score"

As aforementioned, the link prediction models will generate scores to represent the link formation probability, and then classify or rank node pairs by the scores accordingly. Their insight is "the larger link score, the higher probability of link formation", which is followed by all existing methods. However, for two node pairs with the same "link score", do they really have the same link formation probability? If it is not true, all the existing methods can be improved by developing better insight.

For example, we find that some user pairs share the same link scores based on a certain metric like Katz, however, their surrounding topology are totally different, which indicate different link formation probabilities. Essentially, to distinguish these situations, we need to consider more additional information other than only Katz. In other words, the performance of Katz, or any existing methods can be improved if we further examine the links scores with additional information.

In fact, the evolution of the link prediction methods has followed the same pattern which further examines the user pairs with the same link score. For example, to cover the shortage of the metric CN, researchers proposed AA [31], JC [3], PA [32] and RA [33] etc, to normalize CN with additional information to generate fair rankings. However, current methods do not consider the characters of each individual neighbors, which can be served as additional information. One promising

direction is to examine the node pairs with the same score generated by the existing neighbor-based methods, and adopt neighbors' degree distribution as additional information to further improve the prediction performance.

List of Author's Publications

- **Xiaoming Li**, Hui Fang, Jie Zhang, "Supervised User Ranking in signed networks." in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 184–191.

- **Xiaoming Li**, Hui Fang, Jie Zhang, "File: A novel framework for predicting social status in signed networks." in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 330–337.

- **Xiaoming Li**, Hui Fang, Qing Yang, Jie Zhang,"Who is your best friend?: Ranking social network friends according to trust relationship," in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP)*. ACM, 2018, pp. 301–309.

- **Xiaoming Li**, "Towards practical link prediction approaches in signed social networks," in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP)*, ACM, 2018, pp. 269–272.

- **Xiaoming Li**, Hui Fang, Jie Zhang,"A feature-based approach for the redefined link prediction problem in signed networks," in *Proceedings of the International Conference on Advanced Data Mining and Applications (AMDA)*, 2017, pp. 165–179.

- **Xiaoming Li**, Hui Fang, Jie Zhang, "Rethinking the link prediction problem in signed social networks." in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017, pp. 4955–4956.

Bibliography

[1] Adam D Farmer, CEM Bruckner Holt, MJ Cook, and SD Hearing. Social networking sites: a novel portal for communication. *Postgraduate medical journal*, 85(1007):455–459, 2009. 1

[2] Zhepeng Li, Xiao Fang, and Olivia R Liu Sheng. A survey of link recommendation for social networks: Methods, theoretical foundations, and future research directions. *ACM Transactions on Management Information Systems (TMIS)*, 9(1):1–26, 2017. 1

[3] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology*, 58(7):1019–1031, 2007. 1, 13, 14, 28, 35, 55, 58, 90

[4] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011. 1

[5] Jiawei Zhang, Yuanhua Lv, and Philip Yu. Enterprise social link recommendation. In *CIKM*, pages 841–850. ACM, 2015. 1

[6] Tong Zhao, H Vicky Zhao, and Irwin King. Exploiting game theoretic analysis for link recommendation in social networks. In *CIKM*, pages 851–860. ACM, 2015. 1

[7] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*, pages 641–650. ACM, 2010. 3, 4, 9, 16, 23, 26, 27, 28, 34, 36, 69, 88

[8] Kai-Yang Chiang, Nagarajan Natarajan, Ambuj Tewari, and Inderjit S Dhillon. Exploiting longer cycles for link prediction in signed networks. In

Proceedings of the 20th ACM International Conference on Information and knowledge management, pages 1157–1162. ACM, 2011. 16, 28

[9] Cho-Jui Hsieh, Kai-Yang Chiang, and Inderjit S Dhillon. Low rank modeling of signed networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–515. ACM, 2012. 16, 17, 23, 34, 36, 49, 58

[10] Jihang Ye, Hong Cheng, Zhe Zhu, and Minghua Chen. Predicting positive and negative links in signed social networks by transfer learning. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1477–1488. ACM, 2013. 3

[11] Dongjin Song and David A Meyer. Recommending positive links in signed social networks by optimizing a generalized auc. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, pages 290–296, 2015. 3, 6, 17, 22, 36, 58, 61, 78

[12] Athanasios Papaoikonomou, Magdalini Kardara, Konstantinos Tserpes, and Theodora A Varvarigou. Predicting edge signs in social networks using frequent subgraph discovery. *IEEE Internet Computing*, 18(5):36–43, 2014. 4, 16

[13] Arti Patidar, Vinti Agarwal, and KK Bharadwaj. Predicting friends and foes in signed networks using inductive inference and social balance theory. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 384–388. IEEE Computer Society, 2012. 16

[14] Thomas DuBois, Jennifer Golbeck, and Aravind Srinivasan. Predicting trust and distrust in social networks. In *SocialCom*, pages 418–424. IEEE, 2011. 4, 16

[15] Purnamrita Sarkar, Deepayan Chakrabarti, and Andrew W Moore. Theoretical justification of popular link prediction heuristics. In *Proceedings of the Twenty-second International Joint Conference on Artificial Intelligence (IJCAI)*, volume 22, page 2722, 2011. 5, 18

[16] Michael J Brzozowski and Daniel M Romero. Who should i follow? recommending people in directed social networks. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011. 5, 18

[17] Zhijun Yin, Manish Gupta, Tim Weninger, and Jiawei Han. A unified framework for link recommendation using random walks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 152–159. IEEE, 2010. 5, 18

[18] Huan Zhao, Xiaogang Xu, Yangqiu Song, Dik Lun Lee, Zhao Chen, and Han Gao. Ranking users in social networks with higher-order structures. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 330–337, 2018. 5, 18

[19] Tong Man, Huawei Shen, Shenghua Liu, Xiaolong Jin, and Xueqi Cheng. Predict anchor links across social networks via an embedding approach. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, volume 16, pages 1823–1829, 2016. 5, 18

[20] Hao Wang, Xingjian Shi, and Dit-Yan Yeung. Relational deep learning: A deep latent variable model for link prediction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 2688–2694, 2017.

[21] Arun Reddy Nelakurthi and Jingrui He. Finding cut from the same cloth: Cross network link recommendation via joint matrix factorization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 1467–1473, 2017. 5, 18

[22] Panagiotis Symeonidis and Eleftherios Tiakas. Transitive node similarity: predicting and recommending links in signed social networks. *World Wide Web*, 17(4):743–776, 2014. 5, 18, 28, 78

[23] Tianchen Zhu, Zhaohui Peng, Xinghua Wang, and Xiaoguang Hong. Measuring the similarity of nodes in signed social networks with positive and negative links. In *Asia-Pacific Web and Web-Age Information Management Joint Conference on Web and Big Data (APWeb-WIM)*, pages 399–407. Springer, 2017. 5, 18, 78

[24] Moshen Shahriari and Mahdi Jalili. Ranking nodes in signed social networks. *Social Network Analysis and Mining*, 4(1):172, 2014. 5, 18

[25] Zhaoming Wu, Charu C Aggarwal, and Jimeng Sun. The troll-trust model for ranking in signed networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 447–456. ACM, 2016. 5, 18

[26] Xiaoming Li, Hui Fang, and Jie Zhang. A feature-based approach for the redefined link prediction problem in signed networks. In *International Conference on Advanced Data Mining and Applications*, pages 165–179. Springer, 2017. 6, 21, 58, 78

[27] Xiaoming Li, Hui Fang, and Jie Zhang. File: A novel framework for predicting social status in signed networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 330–337, 2018. 6, 43

[28] Xiaoming Li, Hui Fang, and Jie Zhang. Supervised user ranking in signed social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 184–191, 2019. 7, 65

[29] Jinhong Jung, Woojeong Jin, Lee Sael, and U Kang. Personalized ranking in signed networks using signed random walk with restart. In *Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 973–978. IEEE, 2016. 7, 18, 65, 69, 78

[30] Jiliang Tang, Yi Chang, and Huan Liu. Mining social media with social theories: a survey. *ACM SIGKDD Explorations Newsletter*, 15(2):20–29, 2014. 9, 26, 88

[31] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003. 14, 90

[32] Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001. 14, 28, 90

[33] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal B-Condensed Matter and Complex Systems*, 71(4):623–630, 2009. 14, 90

[34] Jianpei Zhang, Yuan Zhang, Hailu Yang, and Jing Yang. A link prediction algorithm based on socialized semi-local information. *Journal of Computational Information Systems*, 10(10):4459–4466, 2014. 14

[35] Weiping Liu and Linyuan Lü. Link prediction based on local random walk. *EPL (Europhysics Letters)*, 89(5):58007, 2010. 14

[36] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002. 14

[37] Yuxiao Dong, Qing Ke, Bai Wang, and Bin Wu. Link prediction based on local information. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 382–386. IEEE, 2011. 14

[38] Elizabeth A Leicht, Petter Holme, and Mark EJ Newman. Vertex similarity in networks. *Physical Review E*, 73(2):026120, 2006. 14

[39] Fei Tan, Yongxiang Xia, and Boyao Zhu. Link prediction in complex networks: a mutual information perspective. *PloS one*, 9(9):e107056, 2014. 14

[40] Zhen Liu, Qian-Ming Zhang, Linyuan Lü, and Tao Zhou. Link prediction in complex networks: A local naïve bayes model. *EPL (Europhysics Letters)*, 96 (4):48007, 2011. 14

[41] Carlo Vittorio Cannistraci, Gregorio Alanis-Lobato, and Timothy Ravasi. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Scientific reports*, 3:1613, 2013. 14

[42] Hon Nian Chua, Wing-Kin Sung, and Limsoon Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006. 14

[43] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953. 14, 28, 35

[44] Ryan N Lichtenwalter and Nitesh V Chawla. Vertex collocation profiles: subgraph counting for link analysis and prediction. In *Proceedings of the 21st International Conference on World Wide Web*, pages 1019–1028. ACM, 2012. 14

[45] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Fast random walk with restart and its applications. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 613–622. IEEE, 2006. 14

[46] Vincent D Blondel, Anahí Gajardo, Maureen Heymans, Pierre Senellart, and Paul Van Dooren. A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM review*, 46(4):647–666, 2004. 14

[47] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 243–252. ACM, 2010. 14

[48] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002. 14

[49] Daniel Spielman. Spectral graph theory. *Lecture Notes, Yale University*, pages 740–0776, 2009. 14

[50] Pavel Chebotarev and Elena Shamis. The matrix-forest theorem and measuring relations in small social groups. *arXiv preprint math/0602070*, 2006. 14

[51] Alexis Papadimitriou, Panagiotis Symeonidis, and Yannis Manolopoulos. Fast and accurate link prediction in social networking systems. *Journal of Systems and Software*, 85(9):2119–2132, 2012. 14

[52] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *Sixth SIAM International Conference on Data Mining: Workshop on Link Analysis, Counter-terrorism and Security*, 2006. 15

[53] Aditya Krishna Menon and Charles Elkan. Link prediction via matrix factorization. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 437–452. Springer, 2011. 15

[54] Jérôme Kunegis and Andreas Lommatzsch. Learning spectral graph transformations for link prediction. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 561–568. ACM, 2009. 15

[55] Sami Abu-El-Haija, Bryan Perozzi, Rami Al-Rfou, and Alexander A Alemi. Watch your step: Learning node embeddings via graph attention. In *Advances in Neural Information Processing Systems*, pages 9180–9190, 2018. 15

[56] Weiwei Gu, Fei Gao, Xiaodan Lou, and Jiang Zhang. Link prediction via graph attention network, 2019. 15

[57] Wentao Wang, Lintao Wu, Ye Huang, Hao Wang, and Rongbo Zhu. Link prediction based on deep convolutional neural network. *Information*, 10(5): 172, 2019. 15

[58] Jérôme Kunegis, Stephan Schmidt, Andreas Lommatzsch, Jürgen Lerner, Ernesto W De Luca, and Sahin Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. In *Proceedings of the 2010 SIAM International conference on Data Mining*, pages 559–570. SIAM, 2010. 15

[59] Priyanka Agrawal, Vikas K Garg, and Ramasuri Narayanam. Link label prediction in signed social networks. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013. 16

[60] Yi Cen, Rentao Gu, and Yuefeng Ji. Sign inference for dynamic signed networks via dictionary learning. *Journal of Applied Mathematics*, 2013, 2013. 17

[61] Dongjin Song, David A Meyer, and Dacheng Tao. Efficient latent link recommendation in signed networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1105–1114. ACM, 2015. 17

[62] Srijan Kumar, Francesca Spezzano, VS Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *Proceedings of the IEEE 16th International Conference on Data Mining*, pages 221–230. IEEE, 2016. 17

[63] Dongjin Song and David A Meyer. Link sign prediction and ranking in signed directed social networks. *Social network analysis and mining*, 5(1):52, 2015. 18

[64] Tibor Antal, Paul L Krapivsky, and Sidney Redner. Social balance on networks: The dynamics of friendship and enmity. *Physica D: Nonlinear Phenomena*, 224(1):130–136, 2006. 26, 69

[65] James A Davis and Samuel Leinhardt. The structure of positive interpersonal relations in small groups. 1967. 27

[66] Armin Falk and Urs Fischbacher. A theory of reciprocity. *Games and economic behavior*, 54(2):293–315, 2006. 27

[67] Zeynep Tufekci. Who acquires friends through social media and why?" rich get richer" versus" seek and ye shall find". In *ICWSM*, 2010. 27

[68] Thin Nguyen, Dinh Q Phung, Brett Adams, and Svetha Venkatesh. Towards discovery of influence and personality traits through social link prediction. In *Fifth International AAAI Conference on Weblogs and Social Media*, pages 566–569, 2011. 43

[69] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–1370. ACM, 2010. 46, 50, 51

[70] Jiliang Tang, Yi Chang, Charu Aggarwal, and Huan Liu. A survey of signed network mining in social media. *ACM Computing Surveys (CSUR)*, 49(3):42, 2016. 46

[71] Steven W Duck and Gordon Craig. Personality similarity and the development of friendship: A longitudinal study. *British Journal of Clinical Psychology*, 17 (3):237–242, 1978. 48

[72] William M Bukowski and Betsy Hoza. Popularity and friendship: Issues in theory, measurement, and outcome. 1989. 48

[73] Xiaoming Li, Hui Fang, and Jie Zhang. Rethinking the link prediction problem in signed social networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4955–4956, 2017. 50, 51

[74] Express.co.uk. The average person has this many friends. `https://goo.gl/bN47rq`, 2017. 55

[75] Steven Mazie. Do you have too many facebook friends? `https://goo.gl/zkLTfe`, 2016. 55

[76] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 452–461. AUAI Press, 2009. 58

[77] Rongjing Xiang, Jennifer Neville, and Monica Rogati. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World Wide Web (WWW)*, pages 981–990. ACM, 2010. 67

[78] Georgios Katsimpras, Dimitrios Vogiatzis, and Georgios Paliouras. Determining influential users with supervised random walks. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*, pages 787–792. ACM, 2015. 67

[79] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web Search and Data Mining (WSDM)*, pages 635–644. ACM, 2011. 68, 76

[80] Olivier Chapelle and S Sathiya Keerthi. Efficient algorithms for ranking with svms. *Information retrieval*, 13(3):201–215, 2010. 71

[81] Jimeng Sun, Huiming Qu, Deepayan Chakrabarti, and Christos Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM)*, pages 8–pp. IEEE, 2005. 72, 74

[82] Ghazaleh Beigi, Jiliang Tang, Suhang Wang, and Huan Liu. Exploiting emotional information for trust/distrust prediction. In *Proceedings of the 2016 SIAM international conference on data mining*, pages 81–89. SIAM, 2016. 89

[83] Sanjeev Dhawan, Kulvinder Singh, and Deepika Sehrawat. Emotion mining techniques in social networking sites. *International Journal of Information & Computation Technology*, 4(12):1145–1153, 2014. 89

[84] Sebastien Razakarivony and Frederic Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016. 89

[85] Liangpei Zhang, Gui-Song Xia, Tianfu Wu, Liang Lin, and Xue Cheng Tai. Deep learning for remote sensing image understanding. *Journal of Sensors*, 2016, 2016. 89

[86] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10. ACM, 2016. 89