Human-centric Computing and Information Sciences

December 2021 | Volume 11

www.hcisjournal.com

KIPS
Korea Information Processing Society

KIPS CSWRG
Korea Information Processing Society
Computer Software Research Group

HCIS | Human-centric Computing and Information Sciences

RESEARCH(Research Manuscript) Open Access

Human-centric Computing and Information Sciences (2021) 11:44
DOI: https://doi.org/10.22967/HCIS.2021.11.044
Received: May 24, 2021; Accepted: October 23, 2021; Published: December 15, 2021

Similarity-based Common Neighbor and Sign Influence Model for Link Prediction in Signed Social Networks

Miaomiao Liu[1,2,*], Jingfeng Guo[3], Jing Chen[3], and Yongsheng Zhang[1,2]

Abstract

Numerous advances have been made in prediction for signed networks. However, few methods can perform simultaneous link and sign prediction. Particularly, for networks with special topologies, the prediction performance of negative links is poor. Moreover, methods based on common neighbors rely heavily on the local structure to achieve a high prediction accuracy. Therefore, a novel model was proposed to realize link and sign prediction and improve the prediction accuracy of negative links. First, the concept of sign influence based on structurally balanced rings was introduced. Subsequently, the similarity of the two nodes based on their first- and second-order common neighbors was defined. Finally, the method was optimized by adjusting the step-size parameters. Experiments were performed on six classic datasets. The prediction accuracy of Epinions, Slashdot, and Wikipedia were 95.1%, 92.5%, and 97.7%, respectively. In the network with a 1:1 ratio of positive and negative edges, the sign prediction accuracy can reach 99.3%. The comparison and results of the recommended top *k* links proved the effectiveness of the proposed algorithm and showed its considerably high link prediction precision and low computational complexity in link and sign prediction for sparse networks and negative link prediction.

Keywords

Signed Networks, Link Prediction, Sign Prediction, Sign Influence, Common Neighbors

1. Introduction

With the rapid development of the Internet of Things, machine learning (ML), and artificial intelligence, many social media platforms have emerged, and large amounts of complex and heterogeneous data have been produced [1]. Characterization and prediction of such data have become a popular research topic in the field of social network analyses [2]. In real networks, entities have positive and negative relations; for example, there are friends and enemies in social fields, supports and oppositions in information fields, and promotions and inhibitions in biological fields. The type of network that uses both positive and negative links is called a signed social network or signed network [3], and it is a crucial branch of social networks. To reveal the relationships between users and the structural evolution mechanism, link prediction has been considered a fundamental research area in signed

***Corresponding Author:** Miaomiao Liu (liumiaomiao82@163.com)
[1] School of Computer and Information Technology, Northeast Petroleum University, Daqing, China
[2] The Key Laboratory for Oil Big Data and Intelligent Analysis of Heilongjiang Province, Daqing, China
[3] College of Information Science and Engineering, Yanshan University, Qinhuangdao, China

networks. Link prediction refers to the prediction of the possibility of establishing unknown links and the prediction of the missing sign type based on the observed network structure [4]. Related studies have been performed in many fields, such as trust evaluation [5], recommendation systems [6], and community detection [7]. Moreover, link prediction presents practical applications in the biological field as it can help in guiding experiments, saving time and cost, and improving accuracy. Tuan et al. [8] proposed some new similarity measures based on fuzzy and neutrosophic environments for link prediction in social networks. These measures are widely used in various domains such as the co-authorship network and protein-interaction systems.

Numerous advances have been made for prediction in signed networks. However, most existing methods can only realize sign prediction of edges but miss their sign attributes. Few methods can perform link prediction as well as sign prediction. Liu et al. [9] realized sign prediction and link prediction in signed networks; however, the accuracy of the prediction for signed networks based on the balance and similarity (PSNBS) algorithm depends on the influence factor of the adjustable step size. In addition, relevant studies [10] on mainstream link prediction methods based on the similarity have shown that when the step size is >3, the computational complexity of the algorithm substantially increases without considerable improvement in the prediction accuracy. In particular, for sparse and signed networks with special topologies, the prediction performance of negative links is sometimes poor. In addition, measuring the similarity between users is vital in link prediction. However, many existing studies use only common neighbors of nodes to measure the similarities, and these methods rely heavily on the local structure of social networks. Thus, achieving a high prediction accuracy is difficult. Based on this, a similarity-based-common-neighbors-and-sign-influence model for link prediction in signed networks is proposed. The main contributions of this paper are as follows:

(1) Considering factors such as the step size and the number of paths, the number and degree of first- and second-order common neighbors, the sign of links, the concept of clustering coefficient of common neighbors and sign influence (SI) based on structurally balanced rings are defined; these can effectively capture their impact on the similarity.

(2) By defining the first- and second-order similarities based on the common neighbor clustering coefficient (CNCC) and SI, respectively, a new similarity criterion of the two nodes is obtained. Then, the dual objectives of link prediction and sign prediction and top *k* recommended links can be realized, which can improve the prediction accuracy while ensuring efficiency.

(3) To reduce the complexity of the algorithm, the sensitivity of adjustable step-size parameters is analyzed considering the effects of paths with different step sizes on the similarity; this can further improve the prediction accuracy of the algorithm.

(4) The correctness and effectiveness of the proposed algorithm are verified for several classical signed network datasets. In addition, comparative experiments show that the method has higher prediction accuracy for both large-scale signed networks with conventional and sparse structure and small networks with special topologies.

2. Related Work

Currently, studies on link prediction of signed networks mainly include methods based on node similarity, matrix decomposition and filling, and ML [11]. Methods based on node similarity are mainly combined with the structural balance theory using the local or global information of the signed network to design similarity indicators. Symeonidis and Tiakas [12] realized link prediction by using the similarity of transmission nodes and the Laplacian clustering algorithm. Based on the sign prediction algorithm, called common neighbor-based prediction (CN-Predict), She and Hu [13] proposed the improved common neighbor-based prediction (ICN-Predict) algorithm by fusing the density of the sign; however, its prediction accuracy for negative links is poor. Papaoikonomou et al. [14] extracted the local features of structurally balanced rings and the frequency of frequent subgraphs to construct features for sign prediction; however, the time complexity is high. Chiang et al. [15] considered the general clustering problem in signed networks and proposed a criterion called balance normalized cut. Thereafter, a sign prediction algorithm called measurement of imbalance (MOI) was proposed on the basis of the spectral analysis, which involved measuring the imbalance of rings with a step length of ≤10. Zhang and Wang [16] studied sign prediction based on the local path index and structural balance theory. Chiang et al. [17] proposed a link prediction algorithm called higher-order circle (HOC) for signed networks based on supervised learning, which involved learning and integrating the structural characteristics of ternary, quaternion, and five-membered rings. Zhu and Ma [18] proposed a highly symmetric quadrilateral structure under the structural balance theory, and then extracted the similarities, dissimilarities, and positive and negative attitude tendencies of the node pairs based on the statistical characteristics of the local structure. Considering the importance of negative links, Sheng et al. [19] proposed a link prediction algorithm based on a hidden space mapping the matrix by integrating structural balance and status theories; their algorithm achieved good results on Epinions and Slashdot datasets. Aiming at the uncertain factors of signed networks, Chen et al. [20] used the set pair theory to realize sign prediction by fusing the deterministic relations and uncertain relations in a network. Javari and Jalili [21] applied the definition of relative similarity between clusters to a collaborative filtering (CF) algorithm and proposed a sign prediction method called CF. The clusters were such that the number of inner-cluster negative and inter-cluster positive links were minimal, and the clusters were socially balanced to the highest possible extent. As the complete structure of large social networks cannot be captured easily, and the process of model learning using deep neural networks is unexplainable and uncontrolled. Liu et al. [22] proposed a novel functional network framework that defined the importance of user attributes through the cloud model and performed aggregate computation in computing neurons to define connections between neurons. Further, they used three-way decisions to process the samples in the boundary for optimizing model performance. Experimental results showed its considerably high link prediction precision. In one study [23], a sign prediction algorithm based on the closed triangle model structure (CTMS) was proposed that can infer unknown relation types; experimental results showed its outstanding performance and substantially low computational complexity. To handle link prediction in incomplete signed networks, a model called unlabeled ties-based link prediction was proposed in one study [24] by utilizing the features of labeled and unlabeled ties. The algorithm extracted four types of features: node-based, structural balance-based, status-based, and latent. Then, it adopted transfer learning (TL) methods to train the model and predict the signs of links, which compensated for the incompleteness of target samples. In general, link prediction algorithms based on node similarity are simple and fast and can achieve high prediction accuracy; however, their prediction effect for sparse networks and negative links is often poor.

The link prediction algorithm based on the matrix mainly transforms signed networks into matrixes and uses a trust propagation model, matrix decomposition, or filling to complete sign prediction. Su and Song [25] proposed a low-rank matrix factorization model in which the signs of out-edges and in-edges of neighbor nodes were introduced as offset information. Shen et al. [26] proposed a framework based on projective non-negative matrix factorization that can realize negative link prediction through unsupervised learning by embedding network structures and user attributes. For drawing signed networks and performing link prediction, Kunegis [27] introduced a matrix factorization (MF) method based on the signed graph Laplacian and the concept of signed resistance distances. The method was evaluated on four signed network datasets. For link prediction in large signed networks, a matrix decomposition model based on an asynchronous distributed random gradient descent algorithm was proposed by Zhang et al. [28]. This algorithm highly reduced the size of parameter space and improved the computational efficiency. In general, the sign prediction model based on matrix processing has high computational complexity and is difficult to evaluate. Therefore, the practical application of this type of method in large networks is limited.

In recent years, scholars have studied the link representation and prediction methods based on convolutional and cyclic neural networks through deep learning mechanisms [29]. However, most scholars have focused on link prediction in traditional social networks, and there are relatively few studies on link and sign prediction in signed networks. To gain insight into how a user creates a positive or negative connection to other users, a new representation named inverse square metric was proposed [30]. It used node properties and distance-based metrics to represent edges and proposed a unified framework for sign and link predictions based on the new representation. Experimental results on a group of large networks proved the effectiveness of the proposed method. Due to insufficient labeled data, a novel sign prediction model [31] for unlabeled signed networks was proposed using branch and bound optimized TL (SP-BBTL), which can yield a global optimal solution and does not require the target domain label information. To resolve the sparsity problem, Nasrazadani et al. [32] proposed a method to predict the sign of links based on clustering and CF methods. Network clustering was performed such that the number of positive links within the clusters and the number of negative links between the clusters were as large as possible. Due to the insufficient sign information of links in signed networks, a sign prediction model called tri-domain relationship pattern was proposed [33]. This model selected an intermediate domain to transfer knowledge from source domains to the target domain. Then, it selected the interference instances and eliminated them to further make the transferable knowledge comprehensive. Finally, it trained the sign classifier using transferable knowledge. Liu et al. [34] proposed an effective model to perform link and sign prediction, which integrated algorithms comprising network embedding, network feature engineering, and an integrated classifier. Experiments showed that the proposed model can offer a powerful methodology for multitask prediction in complex networks. Naderi and Taghiyareh [35] proposed an algorithm to improve trust prediction in weighted sign networks by using local variables. The method predicted the sign of edges accurately by computing the stress of related nodes. Most contemporary methods of inferring the trust relation usually ignore semantic relation and the influence of negative links (e.g., distrust relation). In view of this, relation representation learning through signed graph mutual information maximization was proposed by Jing et al. [36]. This learning incorporated a translation model and positive point-wise mutual information to develop a sign prediction model. However, the algorithm relied on the entity and relation semantic spaces to enhance the relation representations. The aforementioned algorithms can achieve sign prediction; however, these cannot achieve simultaneous link prediction and sign prediction.

Given the advantages and disadvantages of the aforementioned methods, to achieve the dual objectives of link prediction and sign prediction, especially the prediction of negative links and signed networks with special topological structures, a link prediction algorithm is proposed that integrates the clustering coefficient of common neighbors and the concept of SI based on the structural balance rings to define the similarity.

3. Proposed Method

3.1 Theoretical Foundation

The structural balance theory provides a theoretical basis for the analysis of undirected signed networks. It starts with the analysis of the balance of triangles (Fig. 1). According to this theory, in an undirected signed network, if the sign product of all edges of a closed ring is positive, then the ring is structurally balanced; otherwise, it is unbalanced. The number of balanced ternary rings in real networks is considerably larger than that of unbalanced rings [37]. The balance index of large signed networks such as Epinions and Slashdot reaches 89.6% and 86.2%, respectively, and the proportion of balanced ternary rings increases with time [38] (Table 1). Some basic laws of the theory are widely used in studies on link prediction in signed networks [27]. Analyzing the sign attributes of the existing links with unknown or missing signs, such as sign prediction, is necessary. Generally, according to the structural balance theory, the ring where two target nodes are located can enhance the structural balance of the network to the maximum extent.

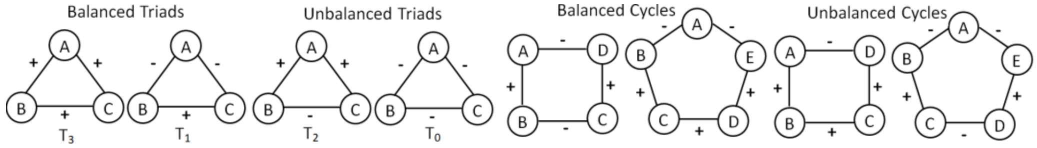


Fig. 1. Structural balance theory.

Table 1. Characteristics of the initial and final signed graph according to the KORIP model [38]

Dataset	Total triangles	Consistent edges		$\|T_0\|$	$\|T_1\|$	$\|T_2\|$	$\|T_3\|$
		Balance	Unbalanced				
Initial GNP	334	8414	2526	41	119	85	103
KORIP GNP	144442865	945810	1960	226500	107963001	678228	35576025
Initial RPL	3218	24698	1947	182	1962	461	627
KORIP RPL	80559455	737350	1981	162684	59991528	492588	19912906

$|T_0|$, $|T_1|$, $|T_2|$, and $|T_3|$ represent the number of triangles of types T_0, T_1, T_2, and T_3 of Fig. 1, respectively.

3.2 Classical Similarity Indicators

For the prediction of signed networks, analysing the possibility of the establishment of links that have not yet been connected, such as link prediction, is necessary. The higher the similarity between two nodes, the higher is the possibility of the establishment of links. Classical similarity indicators include common neighbor (CN), Jaccard, Adamic–Adar (AA), resource allocation (RA), local path (LP), and Katz, as shown in formulae (1)–(6). In these formulae, S_{xy} represents the similarity between nodes x and y, $\Gamma(x)$ represents the neighbor set of node x, $k(x)$ represents the degree of node x, and the symbol "| |" represents the size of the set. α is the parameter for adjusting the influence of the third-order path on the similarity. A is the adjacency matrix of G. The term $path_{xy}^{<l>}$ represents the set of paths with a step size of l between nodes x and y, and β^l represents the damping factor of the path with a step size of l.

$$S_{xy}^{CN} = |\Gamma(x)\cap\Gamma(y)| \tag{1}$$

$$S_{xy}^{Jaccard} = \frac{|\Gamma(x)\cap\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|} \tag{2}$$

$$S_{xy}^{AA} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log k(z)} \tag{3}$$

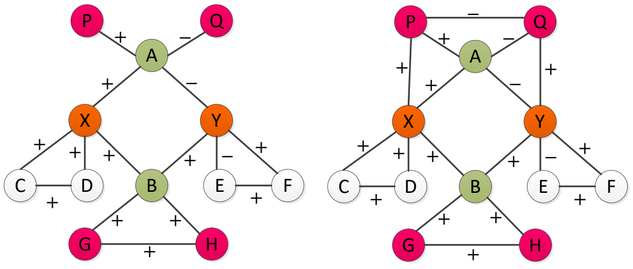
$$S_{xy}^{RA} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{k(z)} \tag{4}$$

$$S_{xy}^{LP} = (A^2)_{xy} + \alpha(A^3)_{xy} \tag{5}$$

$$S_{xy}^{Katz} = \sum_{l=1}^{\infty} \beta^l \times |paths_{xy}^{<l>}| = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + \cdots \tag{6}$$

3.3 Raising Problems

When considering the local topological information of a network, the classical similarity indexes, such as CN, RA, and AA, do not consider the influence of the clustering coefficient of the CNs on the similarity. As shown in Fig. 2, in terms of the node pair <*X*, *Y*>, the degree of the nodes *X* and *Y*, the number of their CNs, and the degree of their CNs in Fig. 2(a) and (b) are the same, and the clustering coefficient of their CN *B* is the same; however, the clustering coefficient of their CN *A* is different. Thus, the concept of CNCC is introduced to comprehensively measure the contribution of characteristics of CNs to their similarity. In addition, the distribution of positive and negative links in signed networks is unbalanced [39], and negative links play a relatively more important role [11, 40]; the number of positive links is considerably larger than that of negative links. Therefore, the concept of SI is introduced to assign different weights to the sign types of multistep paths to highly accurately measure the influence of multiple paths on the sign types.



(a) (b)

Fig. 2. Similarity definition based on the clustering coefficient: (a) Network 1 and (b) Network 2.

In this paper, we propose the CNCC-SI algorithm. To determine the influence of the local-path-information-based ternary ring on the similarity, the CNCC is introduced to comprehensively consider the contribution of the degree of the two nodes, the number of CNs, and their clustering coefficient. To determine the influence of the global-path-information-based balance ring on the similarity of nodes, the concept of SI is introduced to comprehensively evaluate the path connecting the two nodes. Considering the high computational complexity of the path information of high-order step sizes, according to a previous study [41], the path information with 2 and 3 steps is used to define the first- and second-order similarities of the node pair to achieve a better balance between prediction accuracy and calculation efficiency.

3.4 Related Definitions

This paper focuses on link prediction in undirected signed networks. An undirected signed network is usually calculated as G=(V,E,S). $V = \{v_1, v_2, \dots, v_n\}$ represents a node set. $E = \{e(i,j) | v_i, v_j \in V, i \neq j\}$ denotes an edge set. $S = \{sign(i,j) | v_i, v_j \in V, i \neq j\}$ is a sign set. If there is a positive link between two nodes, then $e(i,j) = 1$ and $sign(i,j) = 1$. If there is a negative link between two nodes, then $e(i,j) = 1$ and $sign(i,j) = -1$. If there is no edge between two nodes, then $e(i,j) = 0$ and $sign(i,j) = 0$.

Definition 1 (Clustering coefficient of common neighbors). The clustering coefficient of the common neighbor v_z of the node pair $<v_x,v_y>$ is introduced, denoted as $CNCC^z_{<x,y>}$, as shown in formula (7), where $T^Z_{<x,y>}$ represents the number of connected edges between neighbors of v_z and $k(z)$ represents the degree of v_z.

$$CNCC^z_{<x,y>} = \frac{2 \times T^z_{<x,y>}}{k(z) \times [k(z) - 1]} \tag{7}$$

Definition 2 (Similarity based on first-order common neighbors). To improve the prediction accuracy, the CNCC-SI algorithm considers many factors, such as the degree of the two nodes, the clustering coefficient, the degree of their first-order common neighbors, and the sign of the edge. For all v_x,$v_y \in V$ and $sign(x,y) = 0$, based on the structural balance theory, the similarity between nodes based on first-order common neighbors is defined as $SCN_1 < x,y >$, as shown in formula (8), where $N_1(x)$ and $N_2(x)$ represent the first- and second-order neighbor sets of node v_x, respectively.

$$SCN_1(x,y) = \sum_{|l|=2} SimPath^l_{<x,y>} = \sum_{z \in N_1(x) \cap N_1(y)} \frac{CNCC^z_{<x,y>} \times sign(x,z) \times sign(z,y)}{k(z)} \tag{8}$$

Definition 3 (SI based on balanced rings). In view of the unbalanced proportion of positive and negative links, the concept of SI is introduced in higher-order paths to assign a small weight to negative links and a large weight to positive links, which is recorded as $SIPath_{<x,y>}^{|l|=3}$, as shown in formula (9), where $l = v_x e(v_x, v_p) v_p e(v_p, v_q) v_y$ is the path connecting v_x and v_y with a step size of 3, and v_p and v_q are the two intermediate nodes on the path l, namely $v_p \in N_1(x) \cap N_2(y)$ and $v_q \in N_1(y) \cap N_2(x)$. Parameter α represents the weight of positive links of path *l*, with the value of 1, and β represents the weight of negative links of path *l*, with the value of 0.5.

$$SIPath_{<x,y>}^{|l|=3} = \begin{cases} 3\alpha, & sign(x,p) + sign(p,q) + sign(q,y) = 3 \\ 3\beta, & sign(x,p) + sign(p,q) + sign(q,y) = -3 \\ 2\alpha + \beta, & sign(x,p) + sign(p,q) + sign(q,y) = 1 \\ \alpha + 2\beta, & sign(x,p) + sign(p,q) + sign(q,y) = -1 \end{cases} \tag{9}$$

Definition 4 (Similarity based on second-order common neighbors): Based on the aforementioned definitions, the similarity of the two nodes based on second-order common neighbors is defined through the path information with step size = 3, which is recorded as $SCN_2 < x, y >$, as shown in formula (10).

$$SCN_2(x,y) = \sum_{|l|=3} SimPath_{<x,y>}^{l} = \sum_{|l|=3} \frac{SIPath_{<x,y>}^{|l|=3} \times sign(x,p) \times sign(p,q) \times sign(q,y)}{k(p) + k(q) - 1} \tag{10}$$

Definition 5 (Similarity based on common neighbors): The total similarity between the two unconnected nodes is defined as the sum of similarity scores of the two nodes based on their first- and second-order common neighbors, which is recorded as $SCN(x,y)$, as shown in Eq. (11). $|SCN(x,y)|$ represents the possibility of the node v_x and v_y to establish a link; the sign type of the link is the same as that of $SCN(x,y)$.

$$SCN(x,y) = \sum_{2 \le |l| \le 3} SimPath_{<x,y>}^{l} = SCN_1(x,y) + SCN_2(x,y) \tag{11}$$

3.5 Algorithm Description

Algorithm 1. CNCC_SI

```
Input: $G = (V,E,S)$
Output: SCN($x,y$) and sign($x,y$)
Begin
  Read Dataset File
  For each $v_x, v_y \in V$ do
     If e($x, y$) = 0 or e($x, y$) = 1 $\wedge$ sign($x, y$) = 0 then
        Stack.size = 3; push $x$ in Stack; sNode = $x$;
        While sNode != $y$ and len(Stack) < 3
              Find nNode = sNode.getRelationNodes();
              If nNode in stack : return;
              Else: push nNode in stack; sNode = nNode; Path.append(stack);
        End while
        Repeat the above procedures until all paths have been found
        For each route in Path
            Calculate SCN$_1$($x, y$) → s1 where len(route) = 2
```

```
                Calculate $SCN_2(x, y)$ → s2 where len(route) = 3
                Return s1 + s2 → $SCN(x, y)$
            End for
        End if
        If $SCN(x, y) > 0$, then {$sign(x, y) = 1$} Else {$sign(x, y) = -1$} End if
        Return $sign(x, y)$
    End for
    Sort $|SCN(x,y)|$ in descending order and return top $k$ $<v_x,v_y>$
END
```

4. Experiments and Analysis

4.1 Dataset

Three classical and large-scale real datasets as well as three small datasets that are commonly used in studies on signed networks are used for our experiments, as shown in Table 2. Among these, networks used in clustering reclustering algorithm (CRA) and finding and extracting community (FEC) algorithm are simulation datasets, and Gahuku–Gama sub-tribe (GGS) is the real dataset [42].

Table 2. Basic characteristics of datasets

Dataset	|V|	|E|	Proportion of positive edges (%)	Average degree	Average shortest path	Average clustering coefficient
Epinions	131828	840799	85.0	12.76	3.16	0.19
Slashdot	79120	515397	77.4	13.02	3.57	0.08
Wikipedia	138592	740106	78.7	10.78	4.01	0.07
CRA	36	74	93.2	4	3.53	0.47
FEC	28	42	71.4	3	3.16	0
GGS	16	58	50.0	7.25	1.54	0.54

4.2 Evaluation Metrics

4.2.1 AUC′

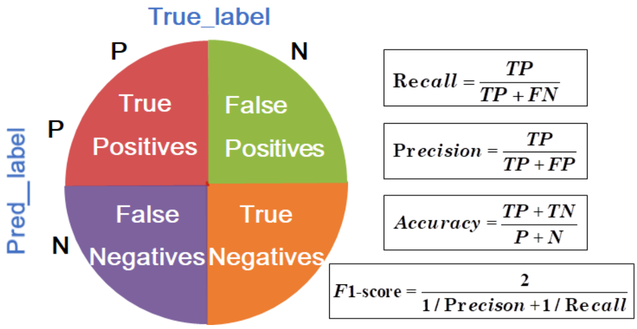
In our algorithm, the total similarity calculated may be positive or negative. Therefore, we adjusted the classic index of the area under the receiver operating characteristic curve (AUC) [43] to obtain a new indicator called AUC′, as shown in formula (12). The probability that a link is selected randomly from the test set has a considerably higher score than a link randomly selected from the non-existent link set. In our experiment, the similarity score corresponding to the randomly selected edge from the test set and the non-existent edge is calculated. Only when the two signs are the same, the scores are compared. If the absolute value of the score of the edge in the test set is larger than that of the non-existent edge, set $\overline{n}' = \overline{n}' + 1$. If the two scores are equal, set $\overline{n}'' = \overline{n}'' + 1$. If the signs of the two are different, edges are selected again. $\overline{n}$ is the total number of experiments in each group in 10 independent experiments, and in this paper, its value is 20,000.

$$AUC' = \frac{\overline{n}' + 0.5 \times \overline{n}''}{\overline{n}} \tag{12}$$

4.2.2 Accuracy′

For sign prediction, TP, FP, TN, FN, Recall, Precision, Accuracy, and F1-score are the commonly used evaluation indicators [4] (Fig. 3). Sign prediction of signed networks must evaluate the comprehensive index of the prediction accuracy of positive and negative links. Relevant studies [4, 19, 26, 39] have shown that the ratio of the number of positive links to negative links in most real networks is more than 4:1, that is, the probability of positive links being selected is substantially higher than that of negative links being selected in the experiment. Therefore, we adjusted the indicators, assigning weight of 1 and 0.5 to the sign prediction results of positive and negative links, respectively. Finally, we use adjusted Accuracy′ to comprehensively evaluate the accuracy of sign prediction, as shown in formula (13).

$$Accuracy' = \frac{TP + 0.5 \times TN}{(TP + FN) + 0.5 \times (TN + FP)} \tag{13}$$



$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{P + N}$$

$$F1\text{-score} = \frac{2}{1/Precison + 1/Recall}$$

Fig. 3. Confusion matrix and common performance metrics.

4.3 Experimental Results and Analysis

Experimental datasets were divided using the ten-fold cross-validation method. The ratio of the training set and test set was 9:1. Further, AUC, AUC′, accuracy, and Accuracy′ were used as evaluation indexes to verify the prediction accuracy of the proposed algorithm.

4.3.1 Link prediction results based on AUC′

Taking AUC′ as the evaluation index, the link prediction accuracies of the proposed and PSNBS [9] algorithms were compared. The average value of 10 independent experiments is shown in Fig. 4. For the first five datasets, PSNBS yielded the highest prediction accuracy under the condition that the step-size influence factor λ is optimal. The proposed algorithm achieved high performance, especially for the CRA network, which has an unbalanced distribution of positive and negative links. Additionally, the proposed algorithm has higher link prediction accuracy than PSNBS. For the GGS network, the accuracy of the proposed algorithm is low. The network describes the political alliances and antagonistic relationships among 16 sub-tribes. Its topology is special (Figs. 5–7). For the dataset with the same number of positive and negative links, the accuracy of the algorithm can still reach 71%, which shows its good robustness.

For the FEC network, the AUC′ of the two algorithms is always 0.5. The topology of the dataset is also very special (Figs. 8, 9). In the FEC network, 28 nodes with a clustering coefficient of 0 are divided into two sets with the same degree distribution, represented by different colours in Fig. 8, where the solid and dotted lines represent positive and negative links, respectively. Among these, 24 nodes have a positive degree of 2 and a negative degree of 1, and the remaining four nodes have a positive degree of 3 and a negative degree of 0. When calculating AUC′, in most cases, the topology of the nodes corresponding to links obtained from the test set and non-existent edge are almost the same, and the probability of their difference is $C_{24}^{2}C_{4}^{2}/C_{28}^{2}C_{26}^{2} = 0.0135$, that is, $\overline{n}' \approx 0$ and $\overline{n}'' \approx \overline{n}$. Therefore, the AUC′ should be 0.5. Experimental results further verify the accuracy of the proposed algorithm.

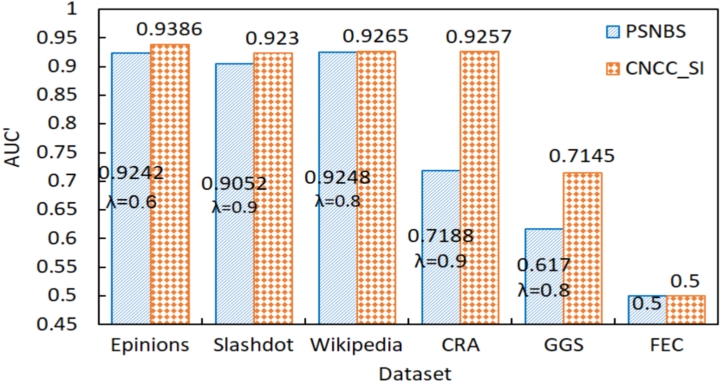


Fig. 4. Link prediction results based on AUC′.

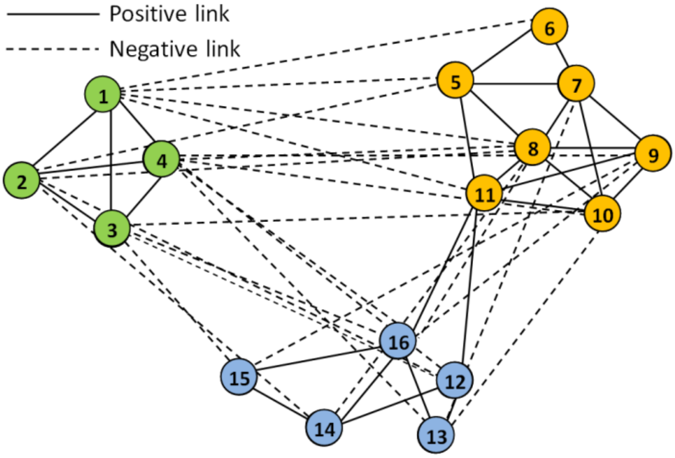


Fig. 5. Topology of GGS network.

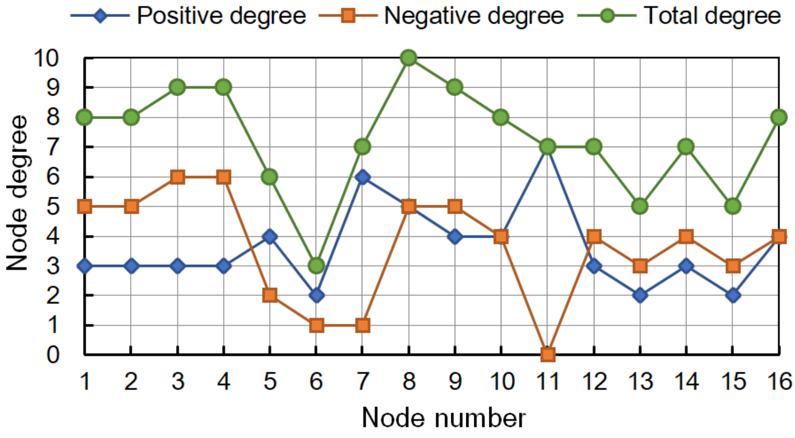


Fig. 6. Degree distribution of GGS.

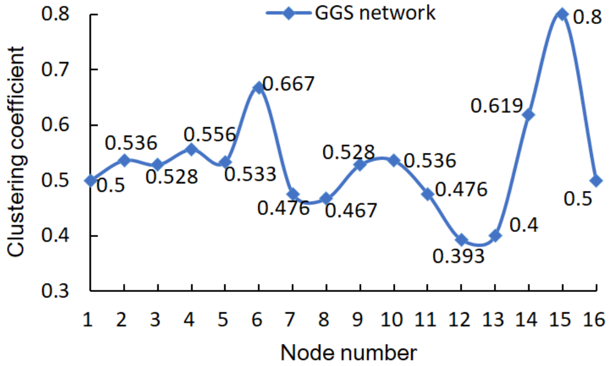


Fig. 7. Clustering coefficient distribution of GGS.

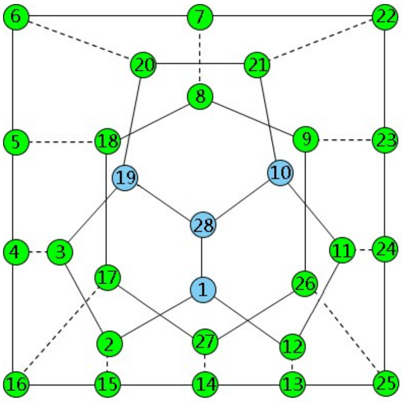


Fig. 8. Topology of FEC.

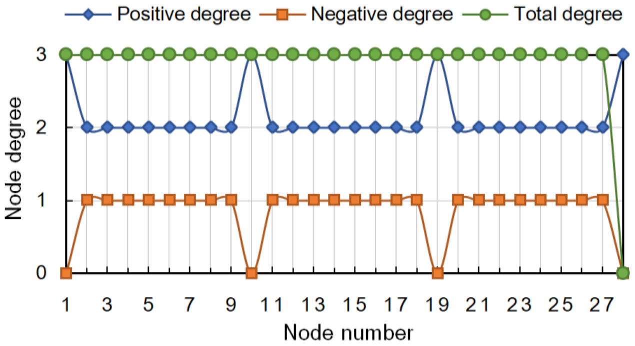


Fig. 9. Degree distribution of FEC.

4.3.2 Sign prediction results based on Accuracy′

Experiments were performed by considering TP, FP, TN, FN, Recall, Precision, F1-score, and Accuracy as evaluation indexes (Table 3). The algorithm exhibited good performance in large networks with conventional topology, small simulation, and real datasets with special topology. It also has high prediction accuracy for negative links, which showed its good robustness.

Table 3. Sign prediction results of CNCC_SI

	Epinions	Slashdot	Wikipedia	CRA	FEC	GGS
TP(+/+)	0.8215	0.6895	0.8581	0.6	0.8750	0.5000
FP(+/–)	0.0555	0.1071	0.0635	0.4	0.1250	0.1667

	Epinions	Slashdot	Wikipedia	CRA	FEC	GGS
TN(–/–)	0.1040	0.1445	0.0635	0	0	0.3333
FN(–/+)	0.0190	0.0589	0.0149	0	0	0
Recall	0.9774	0.9213	0.9829	1	1	1
Precision	0.9367	0.8655	0.9311	0.6	0.8750	0.7500
F1-score	0.9566	0.8926	0.9563	0.75	0.9333	0.8571
Accuracy	0.9255	0.8340	0.9216	0.6	0.8750	0.8333

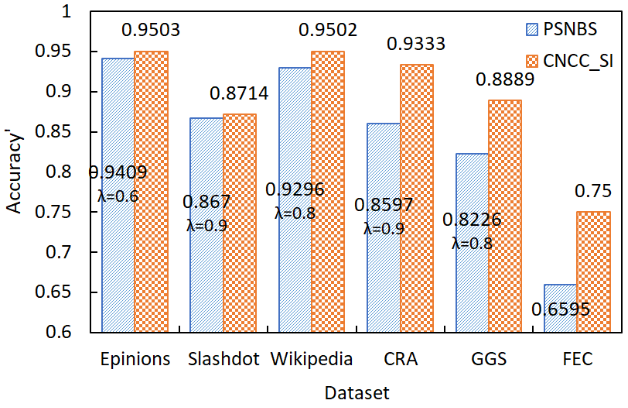


Fig. 10. Prediction results based on Accuracy′.

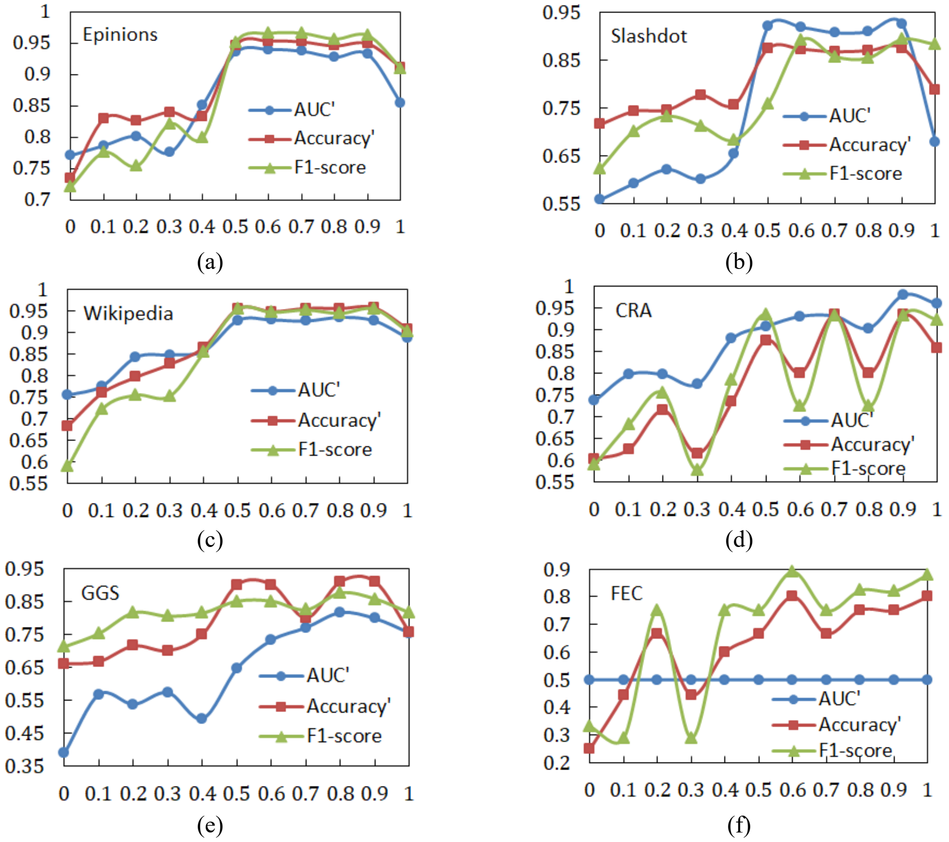
Moreover, considering Accuracy′ as the evaluation index, a comparison between the proposed algorithm and the PSNBS algorithm was performed. The results shown in Fig. 10 are the average value of 10 independent experiments. The sign prediction accuracy of the CNCC-SI algorithm is higher than that of the PSNBS algorithm. Particularly, in the case of the three small networks with special topological structures, the sign prediction accuracy of the proposed algorithm considerably improved, which further showed its correctness and effectiveness.

4.4 Sensitivity Analysis of Adjustable Step Size

Related studies have shown that the contribution of higher-order paths to the similarity of nodes is lower than that of lower-order paths [9, 10]. Thus, in subsequent experiments, the adjustable step parameters ε ($0.5 \leq \varepsilon \leq 1$) and $1 - \varepsilon$ were given to the paths with step sizes of 2 and 3, respectively. The total similarity of the two nodes was modified, as shown in formula (14), which was recorded as $\mathrm{SCN}(x, y)^{\varepsilon}$, and the modified algorithm was marked as $\mathrm{CNCC\_SI}^{\varepsilon}$.

$$\mathrm{SCN}(x,y)^{\varepsilon} = \varepsilon \times \mathrm{SCN}_1(x,y) + (1-\varepsilon) \times \mathrm{SCN}_2(x,y) \qquad (14)$$

Accordingly, experiments were performed under the same conditions. The adjustable step-size parameter ε was between 0 and 1. The experimental results based on AUC′, Accuracy′, and F1-score are shown in Fig. 11. For a particular network, the changing trend of the link prediction and sign prediction with ε is consistently based on AUC′, Accuracy′, and F1-score, which verified the correctness of the proposed algorithm to some extent.



(a) (b)

(c) (d)

(e) (f)

Fig. 11. Link prediction results of CNCC-SI$^{\varepsilon}$ under different ε: (a) Epinions, (b) Slashdot, (c) Wikipedia, (d) CRA, (e) GGS, and (f) FEC.

4.5 Comparison with Other Algorithms

4.5.1 Comparison based on AUC

Experiments were repeated on three large datasets by considering the AUC presented in [13] as the evaluation index. For the AUC index [13], the numbers of positive and negative links predicted correctly were assigned weights of 1 and 0.5, respectively. The results of CN-Predict [13], ICN-Predict [13], PSNBS(λ) [9], CNCC-SI, and CNCC-SI$^{\varepsilon}$ algorithms are shown in Fig.12. The prediction accuracy of the CNCC-SI$^{\varepsilon}$ algorithm is higher than that of the others. The similarity calculation method based on the clustering coefficient of common neighbors and the SI can effectively solve the problems associated with other algorithms, especially for networks with special topologies.

It should be noted that the theoretical basis of link prediction for signed networks is the structural balance theory and status theory, which are applicable to the undirected signed networks and directed signed networks, respectively. Although, it ignores the direction of edges, the structural balance theory is also suitable for the link prediction of directed signed networks. This paper focuses on the research of link prediction in undirected signed networks, so we ignore the direction of edges in the datasets and transform them into undirected signed networks. However, some scholars pay attention to the link prediction in directed signed networks and use the status theory for sign prediction. For example, Yang et al. [44] proposed a deep sign prediction (DSP) method which took both the balance theory and the status theory into account, and used deep learning technology to capture the structure information of the signed networks. Besides, based on the status theory and the number of edge-dependent motifs, Liu et al. [45] proposed a sign prediction algorithm called Motif Family, which was explained by a naive Bayesian model, and achieved good prediction performance in the link prediction of directed signed networks. In order to further verify the performance of our algorithm, we use the AUC metric described earlier [44] to compare the proposed algorithm CNCC_SI and Motif Family algorithm for positive and negative link prediction. The results are shown in Fig. 13.

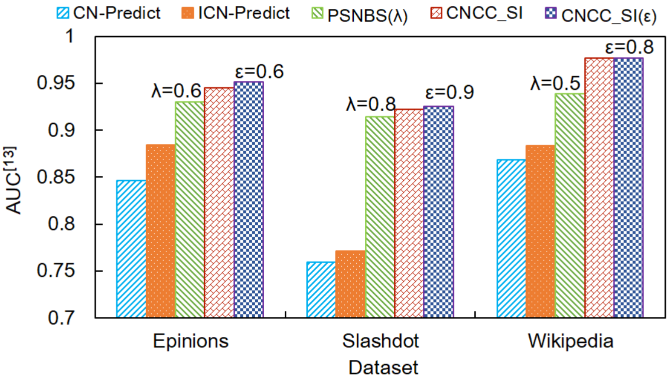


Fig. 12. Comparison based on AUC [13].

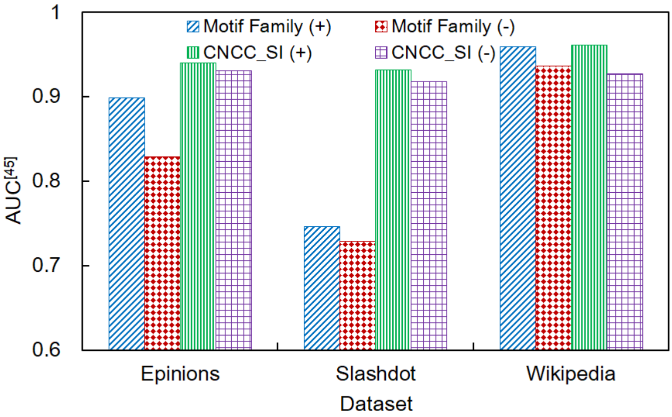


Fig. 13. Comparison based on AUC [45].

The experimental results showed that the AUC values of the prediction results of positive and negative links on the three large datasets of the algorithm CNCC_SI proposed in this paper could achieve satisfactory results. Except for that the AUC result of negative prediction on the Wikipedia was slightly lower than that of the Motif Family algorithm, the positive and negative prediction results of our algorithm on other datasets are much better than that of the Motif Family algorithm. However, the Motif Family algorithm achieved better performance in 16 models for negative prediction. Therefore, for sign prediction in directed signed networks, the future work will be how to combine structural balance theory and status theory effectively to achieve higher prediction accuracy.

4.5.2 Comparison based on Accuracy and F1

Considering Accuracy [4] as the evaluation index, we compared the proposed algorithm with classical algorithms such as MOI [15], HOC [16], CF [21], MF [27], and CTMS [23]. Besides, taking micro-F1 [44] as the metric, we compared our algorithm with the DSP and DSP_B algorithm [44]. The experimental results on large scale datasets are shown in Figs. 14 and 15.

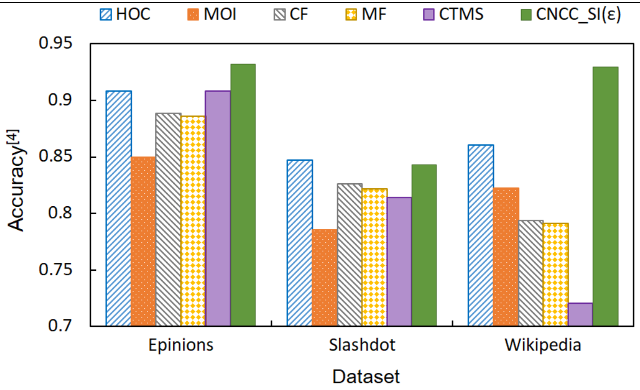


Fig. 14. Comparison of results based on Accuracy [4].

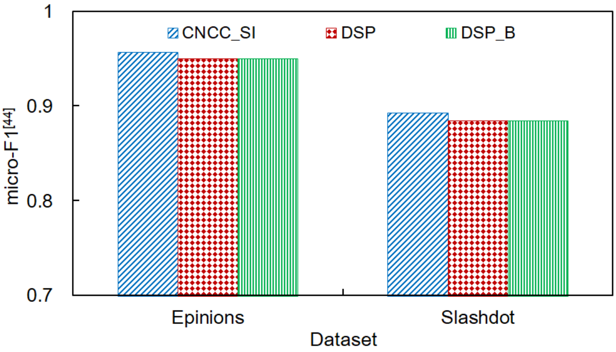


Fig. 15. Comparison based on micro-F1 [44].

The MOI algorithm measured the imbalance of rings with step sizes <10; however, its accuracy was lower than that of other methods. The low accuracy of the CF algorithm also showed that in addition to the structural balance of the network, the local and global structures have a considerable impact on the sign prediction results. The HOC algorithm not only learned the characteristics of ternary rings but also integrated the features of quaternion and five-membered rings. However, its accuracy was worse than that of CNCC-SI, which used only the structural features of ternary and quaternion rings. The aforementioned experimental results showed that the main factor affecting the sign of edges in signed networks is the attribute characteristics of the two endpoints of the edge, followed by the local and global structure characteristics of the connected edge. This finding is consistent with the conclusion in literature [18]. Moreover, the Accuracy [4] of the proposed algorithm for Epinions and Slashdot was 93.2% and 84.3%, respectively. Although it was slightly lower than that of the SPR model (94.7% and 93.8%, respectively) described in [18], the prediction accuracy of the proposed algorithm on Wikipedia (92.9%) was considerably better than that of the SPR algorithm (86.6%). Furthermore, the Accuracy [4] does not distinguish whether the correct sample is positive or negative; thus, the prediction effect of the relevant algorithm is not good for datasets with special topologies. However, the proposed algorithm can achieve the dual goals of link and sign prediction, and it pays relatively more attention to the proportion of positive and negative links and yields high prediction performance for all signed networks. Moreover, when taking the micro-F1 as the metric, the sign prediction result of the CNCC_SI algorithm is slightly higher than that of the DSP algorithm, which shows the superiority of our method.

4.6 Top k Recommended Links

The top 60 links recommended using our algorithm on the three large datasets are shown in Figs. 16–18. These figures show the normalized similarity corresponding to the first three positive links and the first two negative links that are most likely to be established. The label of the node pair corresponding to the link is also marked. In addition, for the two small datasets, recommendation results and the top five links of the PSNBS and the CNCC-SI algorithm were compared in detail (Figs. 19–21); the prediction results of the two algorithms for unknown links are in strong agreement, which further verified the correctness and effectiveness of the proposed algorithm.

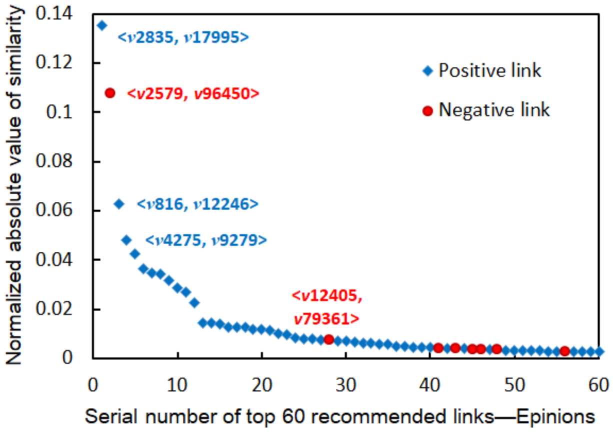


Fig. 16. Recommended links for Epinions.

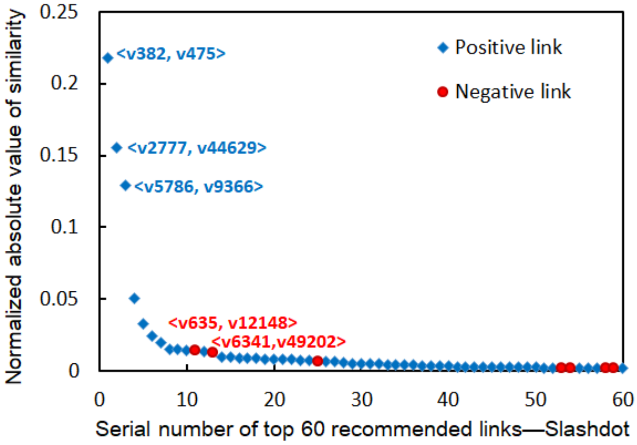


Fig. 17. Recommended links for Slashdot.

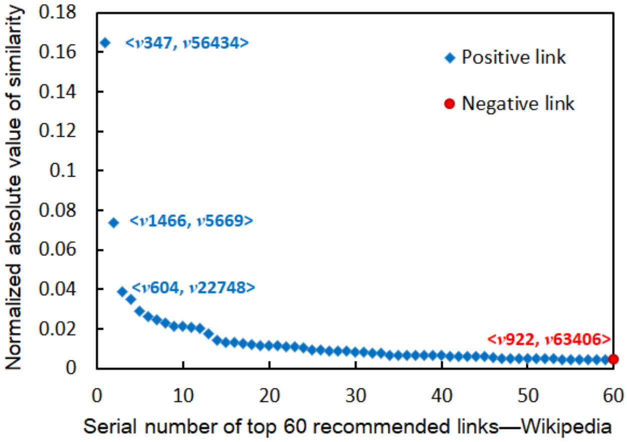


Fig. 18. Recommended links for Wikipedia.

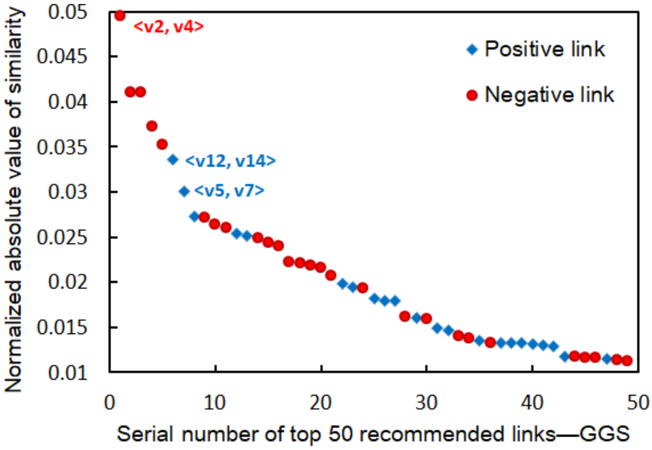


Fig. 19. Recommended links for GGS.

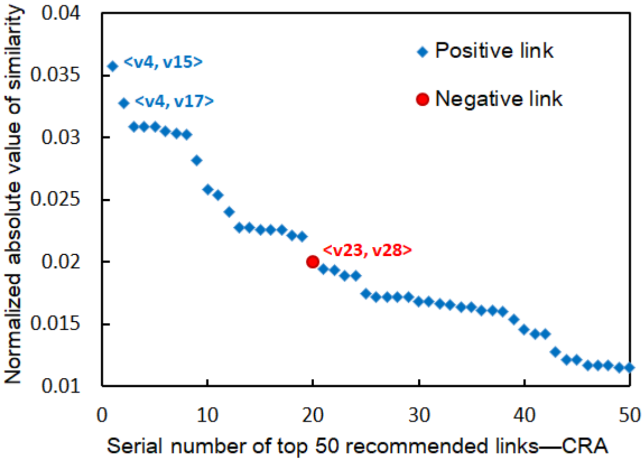


Fig. 20. Recommended links for CRA.

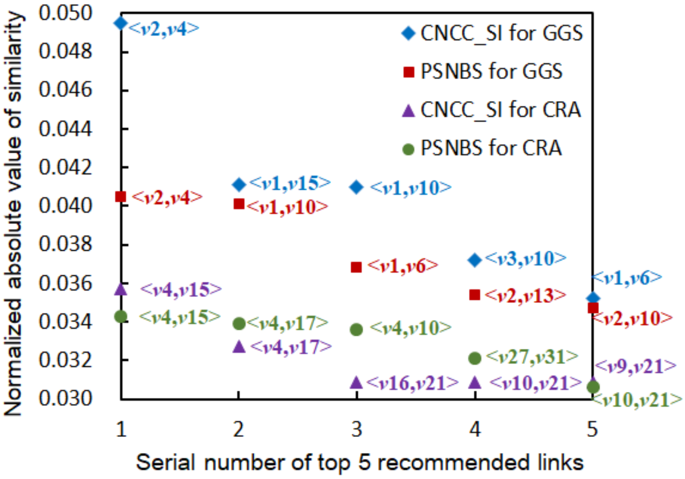


Fig. 21. Comparison of top five recommended links.

5. Conclusion

In this study, the CNCC-SI algorithm is proposed, which can simultaneously predict positive and negative links in signed networks, especially in sparse negative links. The two-step similarity based on first-order common neighbors and the three-step similarity based on second-order common neighbors are defined by combining the clustering coefficient of common neighbor nodes and signed influence. The algorithm is further improved through the sensitivity analysis of the influence factor of the adjustable step-size parameter. The experimental results on several datasets confirm the correctness, high prediction accuracy, and good robustness of the proposed algorithm in link and sign prediction. However, many challenges remain for link prediction in signed networks with complex structures, such as time-dependent networks, multilayer networks, and super-networks. Given the dynamic nature of signed networks, the uncertainty of nodes and relationships, and the data sparseness problem, effective use of the rich information of multidimensional data for fast and accurate link prediction and design of localization or parallel algorithms are the areas for future research.

Acknowledgements

Not applicable.

Author's Contributions

Conceptualization, ML. Funding acquisition, ML, JC. Investigation and methodology, ML. Supervision, JG. Writing of the original draft, ML. Writing of the review and editing, ML, JG, JC, YZ. Formal analysis, YZ.

Funding

This work was supported by the National Natural Science Foundation of China (No. 42002138 and 62172352), Natural Science Foundation of Heilongjiang Province (No. LH2019F042), Youth Science Foundation of Northeast Petroleum University (No. 2018QNQ-01), Excellent Young and Middle-aged Innovative Team Cultivation Foundation of Northeast Petroleum University (No. KYCXTDQ202101).

Competing Interests

The authors declare that they have no competing interests.

References

[1] M. Kim, K. Kim, and J. H. Kim, "Cost modeling for analyzing network performance of IoT protocols in blockchain-based IoT," *Human-centric Computing and Information Sciences*, vol. 11, article no. 7, 2021. https://doi.org/10.22967/HCIS.2021.11.007

[2] R. Mahapatra, S. Samanta, M. Pal, and Q. Xin, "Link prediction in social networks by neutrosophic graph," *International Journal of Computational Intelligence Systems*, vol. 13, no. 1, pp. 1699-1713, 2020.

[3] S. Cheng, H. Shen, G. Zhang, and X. Cheng, "Survey of signed network research," *Journal of Software*, vol. 25, no. 1, pp. 1-15, 2014.

[4] M. Liu, Q. Hu, J. Guo, and J. Chen, "Survey of link prediction algorithms in signed networks," *Computer Science*, vol. 47, no. 2, pp. 21-30, 2020.

[5] G. Wang, J. Park, R. Sandhu, J. Wang, and X. Gui, "Dynamic trust evaluation model based on bidding and multi-attributes for social networks," *International Journal of High Performance Computing and Networking*, vol. 13, no. 4, pp. 436-454, 2019.

[6] L. Zhang, M. Zhao, and D. Zhao, "Bipartite graph link prediction method with homogeneous nodes similarity for music recommendation," *Multimedia Tools and Applications*, vol. 79, no. 19, pp. 13197-13215, 2020.

[7] S. Pulipati, R. Somula, and B. R. Parvathala, “Nature inspired link prediction and community detection algorithms for social networks: a survey,” *International Journal of System Assurance Engineering and Management*, 2021. https://doi.org/10.1007/s13198-021-01125-8

[8] T. M. Tuan, P. M. Chuan, M. Ali, T. T. Ngan, M. Mittal, and L. H. Son, “Fuzzy and neutrosophic modeling for link prediction in social networks,” *Evolving Systems*, vol. 10, no. 4, pp. 629-634, 2019.

[9] M. Liu, J. Guo, and J. Chen, “Link prediction in signed networks based on similarity and structural balance theory,” *Journal of Sichuan University (Engineering Science Edition)*, vol. 50, no. 4, pp. 161-169, 2018.

[10] K. Y. Chiang, N. Natarajan, A. Tewari, and I. S. Dhillon, “Exploiting longer cycles for link prediction in signed networks,” in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, Glasgow, Scotland, UK, 2011, pp. 1157-1162.

[11] M. Lan, C. Li, S. Wang, K. Zhao, Z. Lin, B. Zou, and H. Chen, “Survey of sign prediction algorithms in signed social networks,” *Journal of Computer Research and Development*, vol. 52, no. 2, pp. 410-422, 2015.

[12] P. Symeonidis and E. Tiakas, “Transitive node similarity: predicting and recommending links in signed social networks,” *World Wide Web*, vol. 17, no. 4, pp. 743-776, 2014.

[13] H. She and M. Hu, “Link prediction based on signed networks,” *Journal of Wuhan University of Technology (Information & Management Engineering)*, vol. 37, no. 5, pp. 602-606, 2015.

[14] A. Papaoikonomou, M. Kardara, K. Tserpes, and T. A. Varvarigou, “Predicting edge signs in social networks using frequent subgraph discovery,” *IEEE Internet Computing*, vol. 18, no. 5, pp. 36-43, 2014.

[15] K. Y. Chiang, J. J. Whang, and I. S. Dhillon, “Scalable clustering of signed networks using balance normalized cut,” in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, Maui, HI, 2012, pp. 615-624.

[16] X. Zhang and X. Wang, “Signed network prediction based on structural balance theory and LP algorithm,” *Journal of Yunnan University for Nationalities (Natural Sciences Edition)*, vol. 27, no. 1, pp. 52-57, 2018.

[17] K. Y. Chiang, C. J. Hsieh, N. Natarajan, I. S. Dhillon, and A. Tewari, “Prediction and clustering in signed networks: a local to global perspective,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1177-1213, 2014.

[18] X. Zhu and Y. Ma, “Sign prediction on social networks based nodal features,” *Complexity*, vol. 2020, article no. 4353567, 2020. https://doi.org/10.1155/2020/4353567

[19] J. Sheng, S. Gu, and L. Chen, “Node classification in signed networks based on latent space projection,” *Journal of Computer Applications*, vol. 39, no. 5, pp. 1411-1415, 2019.

[20] X. Chen, J. F. Guo, X. Pan, and C. Zhang, “Link prediction in signed networks based on connection degree,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 5, pp. 1747-1757, 2019.

[21] A. Javari and M. Jalili, “Cluster-based collaborative filtering for sign prediction in social networks with positive and negative links,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 2, pp. 1-19, 2014.

[22] Q. Liu, Y. Chen, G. Zhang, and G. Wang, “A novel functional network based on three-way decision for link prediction in signed social networksm” *Cognitive Computation*, 2021. https://doi.org/10.1007/s12559-021-09873-2

[23] A. Khodadadi and M. Jalili, “Sign prediction in social networks based on tendency rate of equivalent micro-structures,” *Neurocomputing*, vol. 257, pp. 175-184, 2017.

[24] D. Li, D. Shen, Y. Kou, Y. Shao, T. Nie, and R. Mao, “Exploiting unlabeled ties for link prediction in incomplete signed networks,” in *Proceedings of 2019 Third IEEE International Conference on Robotic Computing (IRC)*, Naples, Italy, pp. 538-543.

[25] X. Su and Y. Song, “Local labeling features and a prediction method for a signed network,” *CAAI Transactions on Intelligent Systems*, vol. 13, no. 3, pp. 437-444, 2018.

[26] P. Shen, S. Liu, Y. Wang, and L. Han, “Unsupervised negative link prediction in signed social networks,” *Mathematical Problems in Engineering*, vol. 2019, article no. 7348301, 2019. https://doi.org/10.1155/2019/7348301

[27] J. Kunegis, “Applications of structural balance in signed social networks,” 2014 [Online]. https://arxiv.org/abs/1402.6865.

[28] H. Zhang, G. Wu, and Q. Ling, “Distributed stochastic gradient descent for link prediction in signed social networks,” *EURASIP Journal on Advances in Signal Processing*, 2019, article no. 3, 2019. https://doi.org/10.1186/s13634-019-0601-0

[29] T. Tuna, A. Beke, and T. Kumbasar, “Deep learning frameworks to learn prediction and simulation focused control system models,” *Applied Intelligence*, 2021. https://doi.org/10.1007/s10489-021-02416-0

[30] M. Ahmadalinezhad and M. Makrehchi, “Edge-centric multi-view network representation for link mining in signed social networks,” *Expert Systems with Applications*, vol. 170, article no. 114552, 2021.

[31] W. Yuan, J. Pang, D. Guan, Y. Tian, A. Al-Dhelaan, and M. Al-Dhelaan, “Sign prediction on unlabeled social networks using branch and bound optimized transfer learning,” *Complexity*, vol. 2019, article no. 4906903, 2019. https://doi.org/10.1155/2019/4906903

[32] M. Nasrazadani, A. Fatemi, and M. Nematbakhsh, “Sign prediction in sparse social networks using clustering and collaborative filtering,” *The Journal of Supercomputing*, 2021. https://doi.org/10.1007/s11227-021-03902-5

[33] J. Pang, W. Yuan, and D. Guan, “Tri-domain pattern preserving sign prediction for signed networks,” *Neurocomputing*, vol. 421, pp. 234-243, 2021.

[34] C. Liu, S. Yu, Y. Huang, and Z. K. Zhang, “Effective model integration algorithm for improving link and sign prediction in complex networks,” *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 3, pp. 2613-2624, 2021.

[35] P. T. Naderi and F. Taghiyareh, “Strup: stress-based trust prediction in weighted sign networks,” *SN Computer Science*, vol. 2, article no. 8, 2021. https://doi.org/10.1007/s42979-020-00388-5

[36] Y. Jing, H. Wang, K. Shao, and X. Huo, “Relation representation learning via signed graph mutual information maximization for trust prediction,” *Symmetry*, vol. 13, no. 1, article no. 115, 2021. https://doi.org/10.3390/sym13010115

[37] A. Hassan, A. Abu-Jbara, and D. Radev, “Extracting signed social networks from text,” in *Proceedings of the 2012 Workshop on Graph-Based Methods for Natural Language Processing*, Jeju, Korea, 2012, pp. 6-14.

[38] M. Malekzadeh, M. Fazli, P. J. Khalilabadi, H. Rabiee, and M. Safari, “Social balance and signed network formation games,” in *Proceedings of the 5th International Workshop on Social Network Mining and Analysis (SNAKDD)*, San Diego, CA, 2011.

[39] J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Signed networks in social media,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Atlanta, GA, 2010, pp. 1361-1370.

[40] H. Wang, Q. Yang, L. Fang, and W. Lei, “Maximizing positive influence in signed social networks,” in *Cloud Computing and Security*. Cham, Switzerland: Springer, 2015, pp. 356-367.

[41] S. Teng, Q. Su, D. Liu, and W. Zhang, “Research about optimizing prediction accuracy and time complexity in signed networks,” *Industrial Engineering Journal*, vol. 20, no. 1, pp. 59-64, 2017.

[42] M. M. Liu, Q. C. Hu, J. F. Guo, and J. Chen, “Link prediction algorithm for signed social networks based on local and global tightness,” *Journal of Information Processing Systems*, vol. 17, no. 2, pp. 213-226, 2021.

[43] H. Wang and Z. Le, “Seven-layer model in complex networks link prediction: a survey,” *Sensors*, vol. 20, no. 22, article no. 6560, 2020. https://doi.org/10.3390/s20226560

[44] W. Yang, Y. Wang, and X. Li, “DSP: deep sign prediction in signed social networks,” in *Database Systems for Advanced Applications*. Cham, Switzerland: Springer, 2020, pp. 641-649.

[45] S. Y. Liu, J. Xiao, and X. K. Xu, “Link prediction in signed social networks: from status theory to motif families,” *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1724-1735, 2020.