Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **4**, July **2017**

Link Prediction in Signed Networks based on The Similarity and Structural Balance Theory

Miao-Miao Liu

Northeast Petroleum University
Daqing, Heilongjiang 163318, China
liumiaomiao82@163.com

Jing-Feng Guo and Jing Chen

College of Information Science and Engineering
Yanshan University
Qinhuangdao, Hebei 066004, China

Received March, 2017; revised April, 2017

Abstract. *Most link prediction algorithms based on similarity only consider the local attributes of the node or path structure information, which leads to the difficulty of equilibrium in accuracy and complexity. Additionally, existing algorithms for link prediction in signed networks can only predict the sign of the edge. Hence, a new method PSN_BS(Prediction in Signed Networks based on Balance and Similarity) is proposed which can achieve link prediction and sign prediction simultaneously in signed networks. Firstly, the 2-step and 3-step similarity of the two nodes based on structural balance theory are defined through combining attribute similarity with path similarity on the basis of the choice of optimal step length. Secondly, the total similarity of the two nodes is defined by introducing the step length connectivity factor which is further determined in the later experiment to achieve higher prediction accuracy. Lastly, link and sign prediction are completed according to the total similarity and negative density of the two nodes. Experiments are done on many signed networks using AUC(Area Under the Curve) and precision as evaluation indices, which show the effectiveness and the higher accuracy of the algorithm proposed. Moreover, PSN_BS is superior to CN(Common Neighbor) and ICN(Improved Common Neighbor) algorithm in sign prediction.*
Keywords: Link prediction; Similarity; Signed networks; Structural balance theory

1. **Introduction.** In social networks like Epinions.com and Slashdot.com, links can be divided into two types: the positive link and the negative link where the former represents positive relationships like friends, support, love, etc. and the latter represents negative relationships like enemies, opposition, hate, etc. We use the sign "+" and "-" to mark these relationships respectively and then it brings the emergence of the signed network which refers to the social network in which edges have positive or negative sign attribute[1]. Signed networks exist in many fields in our real world. For example, there are relationships of friends or enemies between users in social networks, and relationships between neurons in biological networks are promotion or inhibition.

The research of signed networks mainly focuses on the analysis of its structure and evolution. One of the hottest issues is link prediction which refers to estimating the possibility of the establishment of a link between two nodes based on the analysis of the network topology[2]. The research of link prediction has a wide range of theoretical value

and practical significance in recommendation systems, attitude prediction and biological fields, etc. Before the negative relationship was introduced into the research of the trust propagation model, researches mainly focused on traditional models that only contain positive relationships, the research on link prediction of signed networks was relatively less. However, related works have shown that researches on negative relationships are always important[3]. The study on link prediction in signed networks can help to analyze the interaction of positive and negative links and promote the application of social networks in many fields such as personalized recommendation, identification of abnormal users and prediction of user behaviors[4]. Existing algorithms for link prediction in signed networks all assume that the edge set of the network are constant which can only achieve sign prediction of the edge where its link type is unknown. However, the real meaning of link prediction in signed networks includes two aspects: link prediction and sign prediction. Moreover, mainstream link prediction algorithms single use local information or path structure of the node to define the similarity, which leads to the deficiency in prediction accuracy. In view of these, an algorithm PSN_BS (Prediction in Signed Networks based on Balance and Similarity) is proposed which effectively integrates the local information and path structure to define the similarity based on structural balance theory. It improves deficiencies of existing algorithms and can achieve link and sign prediction simultaneously. Experiments have also showed its effectiveness and the higher prediction accuracy.

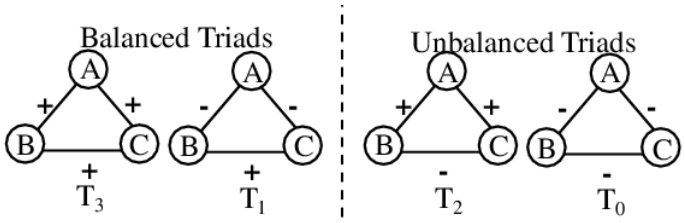

FIGURE 1. Diagrams of balanced and unbalanced triangles

2. **Preliminary.** Structural balance theory was proposed by Heider in 1946 which provided the foundation for structure analysis of signed networks. It used a balanced model to describe the structure derived from positive and negative relationships and conflicts between these links. Then Cartwright and Harary applied the theory into the graph and used an mathematical language to formulate this model as the signed network. Leskovec[5] firstly applied structural balance theory into link prediction of signed networks in 2010.

2.1. **Structural balanced triangles.** Structural balance theory is based on the balance analysis of triangles, which takes all possible configuration models of triads into account. In the undirected signed network, there are totally four configurations denoted by T_0, T_1, T_2, T_3 respectively, as shown in figure 1, where T_i means there are i positive links in this model. Thus four intuitive understandings formed: (1)The friend of my friend is my friend. (2)The enemy of my friend is my enemy. (3)The friend of my enemy is my enemy. (4)The enemy of my enemy is my friend. According to structural balance theory, the balance of a triangle depends on the product of the sign of its three edges. If the product is positive, the triangle is balanced, otherwise it is unbalanced. As described above, T_1 and T_3 are structural balanced while T_0 and T_2 are unbalanced.

2.2. **Structural balanced circles.** The method for judgment of the balance of a triangle can be extended into the balance analysis of circles. A circle is structural balanced if and only if the product of signs of all its edges is positive, as shown in figure 2.

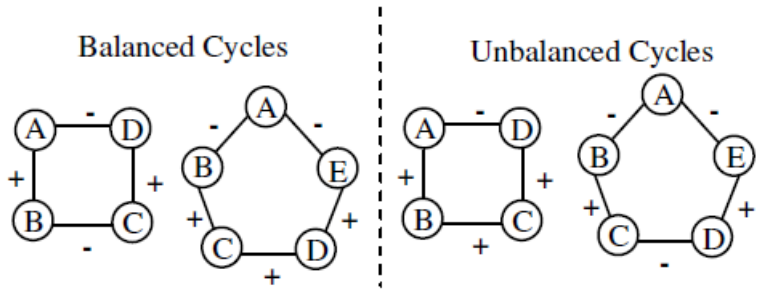

FIGURE 2. Diagrams of balanced and unbalanced circles

2.3. **Balance analysis of signed networks.** If all circles in an undirected signed network are balanced, we say the network is balanced. Pranay[6] analyzed the balance of Epinions and Slashdot and the results were shown in Table 1. Studies of Hassan[7] showed that in signed networks that were simulated or constructed automatically, T_3 model was always fully expressed, while T_0 and T_2 were not fully expressed. Meanwhile the constructed signed network was similar to the clearly expressed signed network and both of them were consistent with structural balance theory. Moreover, related researches in reference [8-10] also found that the number of balanced triangles is far more than unbalanced triangles in real signed networks. And the unbalanced network would evolve towards balanced network as time goes on, which further verifies the importance of structural balance theory in measurement of the balance of undirected signed networks.

TABLE 1. Analysis of global balance of signed networks

Dataset	$\|V\|$	$\|E\|$	$\|T_0\|$	$\|T_1\|$	$\|T_2\|$	$\|T_3\|$	Balance index
Epinions	131828	379603	58732	396548	451711	4003085	90.8%
Slashdot	81871	214996	12075	76859	66679	414956	88.3%

3. **Related work.** Existing link prediction algorithms for signed networks are mainly divided into two categories: sign prediction based on the matrix and sign prediction based on the classification[4].

(1)Sign prediction methods based on the matrix: They convert the signed network into a matrix and use the trust propagation model, matrix factorization or matrix filling to predict the sign of the edge. Priyanka[11] pointed out using the singular value decomposition, eigenvalue decomposition or kernel function decomposition of the matrix can all effectively predict the sign of the edge. Hsieh[12] converted the sign prediction problem into the filling of low rank matrix, which effectively predicted the unknown sign of edges using MC-SVP(Matrix Completion-Singular Value Projection) algorithm. Additionally, some sign prediction algorithms based on the matrix converted the continuous numbers of the matrix of prediction results into the discrete values 1 or -1 through threshold so as to get the final sign prediction result.

(2)Sign prediction methods based on classification: They regard the link prediction in signed networks as the problem of binary classification, which construct feature sets firstly and then complete sign prediction using classification algorithms. According to the difference of information used in construction of the feature set, these algorithms can be divided into two categories: the algorithm based on network structure information and the algorithm based on network context information. The former designed sign prediction algorithms using network structure information on the basis of structural balance theory,

while the latter used context information of the signed network to construct feature set and then predicted the sign through classification algorithms. Leskovec[5] analyzed the difference between link models of real signed networks and link models that were predicted, and the evolution of undirected signed networks was also studied in the paper. Jure[13] completed sign prediction of the underlying network through supervised machine learning using datasets of Epinions, Slashdot and Wikipedia. The algorithm had the higher accuracy but it only considered the local information of the network. Chiang[14] extracted circles with step length of k from the network to build the feature set based on Katz index, then it used the logic regression model to predict the sign of the link. Experiments showed that the accuracy was improved with the increase of k from 3 to 5. When k>5 it changed little. The advantage of this algorithm was it took the relationship between the overall network structure and the sign of the edge into account. Ye[15] accomplished sign prediction through structure knowledge transfer between large scale networks. Borzymek[16] constructed feature set using attributes of the graph such as the degree of the node and attributes of the review such as the number of the comment. Then it used C4.5 decision tree algorithm to train the classifier and completed the sign prediction. Yang[17] proposed a supervised and semi-supervised sign prediction model on the basis of the study of relationships between users and their behaviors. The model was further improved based on structural balance theory and then the prediction accuracy was increased. Facchetti[18] analyzed the structure of large scale signed networks and computed its global balance index, which got the conclusion that the majority of online networks were structural balanced. Panagiotis[19] defined the similarity of basic nodes and transitive nodes to capture local and global features of the network respectively, and completed link prediction based on information of edges. Patidar[20] put forward an inductive learning framework on the basis of structural balance theory and predicted links of friends or enemies using C4.5 algorithm. SHE[21] proposed an improved algorithm ICN(Improved Common Neighbor) based on common neighbors, which completed sign predicted on the basis of combining node density with network topology so as to improve the sign prediction accuracy of negative links.

Above algorithms can only complete sign prediction of existing edges and related researches[22] have found that features based on the network structure are more important than those based on the context of the network in terms of mainstream link prediction algorithms based on similarity. Moreover, a large number of practical analyses have fully verified the effectiveness of structural balance theory in describing the interaction between positive and negative links, the formation principle and modeling of dynamic evolution of the signed network. So based on instantaneous snapshots and topology structure of the signed network, through effectively combining the local information such as node degree and global features such as path structure, the similarity of the two nodes based on structural balance theory is defined and the algorithm PSN_BS is put forward which can achieve link prediction and sign prediction simultanously. The main ideas of the algorithm are described in the fourth part of the paper. The fifth part is definition and description of the algorithm. The last two parts are experiments and the conclusion.

4. **Main ideas.** The goal of PSN_BS is accurate prediction of future links and sign prediction of existing links. The important hypothesis of the algorithm is that the higher the similarity of the two nodes is, the larger possibility they would have to establish a link in the future. In view of the influence of local features and global information of the network to the similarity, similarity contributions of all paths that connect the two nodes are taken into account. We think that the more paths there are between two nodes, the higher similarity they have. And the contribution of the shorter path to the

similarity is greater compared with the longer path. So the algorithm firstly extracted the L-steplength paths that connected the two nodes and then defined the similarity based on the integration of node information and path structure. Meanwhile, considering the effect of future links on structure evolution and balance of the network, the definition of the similarity measurement should also select the network attributes which can reflect the formation mechanism of signed networks from various angles on the basis of structural balance theory so as to accurately predict the sign of the link.

Meanwhile, Zhang[23] found that the structural balance theory of triangles can provide important support for sign prediction, and the prediction accuracy was improved after introducing balanced quadrilaterals compared with single balanced triangles. This further validated that increasing the length of the balanced cicles appropriatly can provide more information for the sign prediction. Additionally, the determination of the optimal step length in STNMP[24](Similarity of Transmission Nodes of Multiple Paths) algorithm proposed by authors of this paper also showed that when using paths of 3 step length to computing the similarity, the complexity of the algorithm was considerably increased, while the prediction accuracy was not significantly improved. In view of these, in this paper we only consider the influence of 2-step and 3-step paths to the similarity of the two nodes so as to reduce the complexity. And different step length connection factors are given to these two types of paths to describe their different contributions to the node similarity. In the end, we use the total contributions of these paths to measure the similarity of the two nodes. In order to accurately describe the algorithm, the definitions of relevant variables are given in table 2.

Table 2. Definition and description of relevant variables

Notation	Implication
G=(V,E,S)	Graph of the undirected signed network
V	Node set. V={ v_1,v_2,$\cdots$,v_n }, $\|V\|$=n
E	Edge set. E={e(v_i,v_j)|e(v_i,v_j) $\in$ {0, 1} }. $\|E\|$=m. $\forall$ v_i,v_j $\in V$ & i$\neq$j, e(v_i,v_j)=e(v_j,v_i), e(v_i,v_j) $\notin E$. If e(v_i,v_j)$\in E$, e(v_i,v_j)=1 else e(v_i,v_j)=0
S	Sign set of edges. S={s(v_i,v_j) }. $\forall$ v_i,v_j $\in V$, s(v_i,v_j) $\in$ {0, 1, -1}. If the link connecting v_i and v_j is positive, s(v_i,v_j)=1. If the link is negative, s(v_i,v_j)=-1. If the link is nonexistent or the sign is unknown, s(v_i,v_j)=0.
$k^+(v_i)$	The positive degree of v_i, namely the number of positive edges that connected with v_i
$k^-(v_i)$	The negative degree of v_i, namely the number of negative edges that connected with v_i
k(v_i)	The degree of v_i, namely the number of edges that connected with v_i. k(v_i)= $k^+(v_i)$+ $k^-(v_i)$
$N_1(v_i)$	The set of the first order neighbor nodes of v_i
$N_2(v_i)$	The set of the second order neighbors of v_i
$l_k(v_i,v_j)$	The No.k path connecting v_i and v_j abbreviated as l_k. Here, $\|l_k(v_i, v_j)\|$ represents the step length of l_k
$S_{lk}(v_i,v_j)$	Sign prediction result of <v_i,v_j> based on l_k
$BScore_2(v_i,v_j)$	2-step similarity score of <v_i,v_j>
$BScore_3(v_i,v_j)$	3-step similarity score of <v_i,v_j>
BScore(v_i,v_j)	Total similarity of <v_i,v_j> based on structural balance theory
BSim	Similarity Matrix of G where BSim(i,j)=BScore(v_i,v_j)

The goal of PSN_BS algorithm can be described as predicting the possibility of the establishment of e(v_i,v_j) and the sign of e(v_i,v_j) on the condition of v_i,v_j $\in V$ and s(v_i,v_j)=0. As shown in figure 3, the solid line represents the existing edge and the dotted line represents the edge that is nonexistent yet. The goal of the algorithm is to compute the possibility of establishing a link between v_1 and v_2 and predict the sign of e(v_1,v_2).

Firstly, in the measurement of the influence of the new link to the balance of the signed network, we define the effect of the path connecting the two nodes on the sign of the edge as the product of the sign of all edges on this path. That is to say, if v_i,v_j $\in V$ & e(v_i,v_j)$\notin E$,

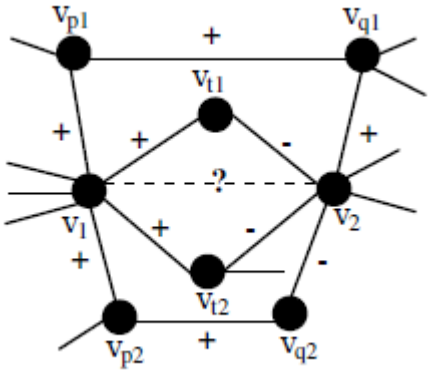

FIGURE 3. A sketch map of PSN_BS algorithm

and $l_k(v_i,v_j)$=v_ie(v_i,v_{k1})v_{k1}e(v_{k1},v_{k2})$v_{k2}\cdots$e(v_{kn},v_j)v_j, then the sign prediction result of e(v_i,v_j) based on l_k is $S_{l_k}(v_i,v_j)$=s(v_i,v_{k1})*s(v_{k1},v_{k2})*$\cdots$*s(v_{kn},v_j). As shown in figure 3, there are four paths connecting v_1 and v_2 which consist of <v_1, v_{t1}, v_2>, <v_1, v_{t2}, v_2>, <v_1, v_{p1}, v_{q1}, v_2> and <v_1, v_{p2}, v_{q2}, v_2> respectively. We denote them as l_1, l_2, l_3 and l_4. Here, $l_1(v_1,v_2)$=v_1e(v_1,v_{t1})v_{t1}e(v_{t1},v_2)v_2 and $S_{l_k}(v_1,v_2)$=s(v_1,v_{t1})*s(v_{t1},v_2)=(+1)*(-1)=-1 . The algorithm should consider the overall impact of these four paths on the sign prediction result of e(v_1,v_2).

Secondly, in the measurement of the effect of own properties of the node on the similarity, we think that the common neighbor with smaller degree contributes more to the similarity of the two nodes than the common neighbor with larger degree. Additionally, among all edges connecting with the common neighbor, there are two edges playing a role in the establishment of the link. Based on these views, the 2-step similarity score of the node pair is defined to measure the similarity contribution of the common neighbor of the path that connects the two nodes with step length of 2. As shown in figure 3, $v_{t1} \in N_1(v_1) \cap N_1(v_2)$, so the similarity contribution of v_{t1} of l_1 to <v_1,v_2> is 2/k(v_{t1}).

Thirdly, in the measurement of the effect of the path structure on the similarity of the node pair, we think that the farther the two nodes are, the less possibility they have to establish a link. In terms of most social networks, the step length of the shortest path is between 3 and 4. In view of this, the 3-step similarity score of the node pair is defined to measure the similarity contributions of paths that connect the two nodes with step length of 3, first order neighbors and second order neighbors of these paths. As shown in figure 3, $v_{p1} \in N_1(v_1) \cap N_2(v_2)$ & $v_{q1} \in N_1(v_2) \cap N_2(v_1)$, then the similarity contribution of v_{p1} and v_{q1} of l_3 to (v_1,v_2) is 3/(k(v_{p1})+k(v_{q1})-1).

Finally, considering the different influence of paths with different step length to the similarity of the two nodes and the sign of the future link, the step length connectivity factor is introduced as adjustable parameter to measure the different similarity contribution of these two types of paths. In the end, we take the sum of contributions of all these paths as the total prediction score of the two nodes based on similarity and structural balance theory. The absolute value of the score measures the similarity of the two nodes which represents the possibility of the establishment of the link. And the sign of the score represents the sign prediction result of the link.

Here, there is a special situation: if the prediction score of the node pair <v_i,v_j> equals 0, maybe the reason is there are no paths that connect v_i and v_j with step length of 2 and 3, which results in the total similarity is 0 and the possibility of the establishment

of e(v_i,v_j) is also 0. Or it may be because the sum of positive and negative value of the 2-step and 3-step similarity score is 0 which also leads to the failure of sign prediction. In this case, the negative density of the node is introduced to analyze the sign tendency of the two nodes to other nodes in the network. When the negative density of v_i and v_j are all larger than the average negative density of the network, it indicates that these two nodes tend to establish negative links with other nodes. At this time, the sign prediction result of e(v_i,v_j) is negative, orelse it is positive.

In summary, the total prediction score of <v_1,v_2> in figure 3 can be expressed as fallows where λ is the step length connectivity factor and $\lambda \in [0,1]$. We know the sign prediction result of e(v_1,v_2) is negative no matter what value of λ.

$$BScore(v_i,v_j) = \lambda * (\frac{2 * s(v_1,v_{t1}) * s(v_{t1},v_2)}{k(v_{t1})} + \frac{2 * s(v_1,v_{t2}) * s(v_{t2},v_2)}{k(v_{t2})}) +$$

$$(1-\lambda) * (\frac{3 * s(v_1,v_{p1}) * s(v_{p1},v_{q1}) * s(v_{q1},v_2)}{k(v_{p1}) + k(v_{q1}) - 1}) + \frac{3 * s(v_1,v_{p2}) * s(v_{p2},v_{q2}) * s(v_{q2},v_2)}{k(v_{p2}) + k(v_{q2}) - 1})$$

$$= (\frac{-3}{5})\lambda + (\frac{-9}{28})(1-\lambda)$$

5. PSN_BS algorithm.

5.1. Relevant definition.

Definition 1: 2-step similarity score of the node pair. Let G=(V,E,S), $\forall$ v_i,v_j $\in V$ & e(v_i,v_j)=0, the 2-step similarity score of the node pair <v_i,v_j> based on structural balance theory is defined as the sum of similarity contributions of all 2-step paths that connect v_i and v_j. We denote it as $BScore_2$<v_i,v_j> which is shown in formula (1). Here, $v_{tk} \in V \cap N_1(v_i) \cap N_1(v_j)$ & e(v_i,v_{tk})=1 & e(v_{tk},v_j)=1.

$$BScore_2(v_i,v_j) = \sum_{k=1}^{|N_1(v_i) \cap N_1(v_j)|} \frac{2}{k(v_{tk})} * s(v_i,v_{tk}) * s(v_{tk},v_j) \tag{1}$$

Definition 2: 3-step similarity score of the node pair. Let G=(V,E,S), $\forall$ v_i,v_j $\in V$ & e(v_i,v_j)=0, the 3-step similarity score of the node pair <v_i,v_j> based on structural balance theory is defined as the sum of similarity contributions of all 3-step paths that connect v_i and v_j. We denote it as $BScore_3$<v_i,v_j> which is shown in formula (2). Here, l_k=v_ie(v_i,v_{pk})v_{pk}e(v_{pk},v_{qk})v_{qk}e(v_{qk},v_j)v_j represents the No.k path that connects v_i and v_j with the step length of 3. And $v_{pk} \in N_1(v_i) \cap N_2(v_j)$ & $v_{qk} \in N_1(v_j) \cap N_2(v_i)$ & e(v_i,v_{pk})=1 & e(v_{pk},v_{qk})=1 & e(v_{qk},v_j)=1.

$$BScore_3(v_i,v_j) = \sum_{|l_k|=3} \frac{3 * s(v_i,v_{pk}) * s(v_{pk},v_{qk}) * s(v_{qk},v_j)}{k(v_{pk}) + k(v_{qk}) - 1} \tag{2}$$

Definition 3: Prediction Score of the Node Pair. Let G=(V,E,S), $\forall$ v_i,$v_j \in V$ & e(v_i,v_j)=0, the prediction score of the node pair <v_i,v_j> based on similarity and structural balance theory is defined as the sum of 2-step and 3-step similarity scores of the node pair<v_i,v_j> . We denote it as BScore< v_i, v_j > which is shown in formula (3). Here, λ is the step length connectivity factor and $\lambda \in [0,1]$.

$$BScore(v_i,v_j) = \lambda * BScore_2(v_i,v_j) + (1-\lambda) * BScore_3(v_i,v_j) \tag{3}$$

Definition 4: Negative Density of the Node. Let G=(V,E,S), $\forall\ v_i \in V$, the negative density of v_i is defined as the ratio of $k^-(v_i)$ to $k(v_i)$. We denote it as $D^-(v_i)$ which is shown in formula (4).

$$D^-(v_i) = \frac{k^-(v_i)}{k(v_i)} \tag{4}$$

Definition 5: Average Negative Density of the Network. Let G=(V,E,S), the average negative density of the network is defined as the average of negative density of all nodes in the network. We denote it as $D^-(G)$ which is shown in formula (5).

$$D^-(G) = \frac{\sum_{i=1}^{n} D^-(v_i)}{n} \tag{5}$$

5.2. **Description of PSN_BS.**

Input: Adjacency matrix of G where A(i,j)=s(v_i, v_j)
Output: Sign prediction result of the edge or top k links that are most likely to establish.
1: ReadGraphFile
2: Initialize Matrix A
3: for each v_i,$v_j \in V$ do
4: if e(v_i, v_j)=0 or s(v_i, v_j)=0
5: Find all 2-step paths $l_k(v_i, v_j)$ where $|l_k|$=2
6: Calcluate $BScore_2(v_i, v_j)$
7: Update Matrix BS_2, namely BS_2(i,j)= $BScore_2(v_i, v_j)$
8: Find all 3-step $l_k(v_i, v_j)$ where $|l_k|$=3
9: Calcluate $BScore_3(v_i, v_j)$
10: Update Matrix BS_3, namely BS_3(i,j)= $BScore_3(v_i, v_j)$
11: Compute BScore(v_i,v_j) and Get Matrix BSim
12: if BScore(v_i, v_j)=0
13: Calcluate $D^-(v_i)$, $D^-(v_j)$ and $D^-(G)$
14: if $(D^-(v_i) > D^-(G))$ and $(D^-(v_j) > D^-(G))$, s(v_i, v_j)=-1
15: else s(v_i, v_j)=+1
16: else if BScore(v_i, v_j) >0, s(v_i, v_j)=+1
17: else s(v_i, v_j)=-1
18: output s(v_i, v_j) end for
19: Sort $|BSim(v_i, v_j)|$
20: Output top k node pairs $< v_i, v_j >$ and the corresponding BScore(v_i, v_j).

6. **Experiments and analysis.** Firstly, datasets were got from Internet and each of them was divided into the training set and the testing set. Secondly, the improved AUC(Area Under the Curve) and Precision indices were used to evaluate the performance of PSN_BS. Lastly, comparison of PSN_BS with classical sign prediction algorithm CN(Common Neighbor) and ICN[21] were done which showed the higher prediction accuracy of the algorithm proposed.

6.1. **Datasets.** Download three real signed networks from http://snap.stanford.edu/ and get their subsets after data processing and extraction. These three datasets are Epinions, Slashdot and Wikipedia. Among them, the positive and negative links in the first two networks are anonymously expressed, and the third network is clearly expressed. In our research, the direction of the link is ignored and these three datasets are transformed

into undirected signed networks. In addition, another two datasets, namely, the Gahuku-Gama network and an illustrated signed network I used in reference [25] are also used in our experiments. The network topology informations of these datasets are shown in table 3 and details are as follows. Here, $|E^+|$ and $|E^-|$ represents the number of positive links and negative links of the network respectively.

TABLE 3. Topology information of datasets

Dataset	$\|V\|$	$\|E\|$	$\|E^+\|/\|E\|$	$\|E^-\|/\|E\|$
(1) Epinions	131828	840799	85%	15%
(2) Slashdot	79120	515397	77.4%	22.6%
(3) Wikipedia	138592	740106	78.7%	21.3%
(4) Gahuku-Gama	16	58	50%	50%
(5) Illustrated signed network I	28	42	71.4%	28.6%

(1)Epinions: It is a consumer review website. Users of the network can view comments of others and create a directed link to express their trust or distrust in others.

(2)Slashdot: It is a technological blog website. All news was provided by its users which can comment news published on the site. Meanwhile users can add others into their friend or enemy list that was represented by positive links and negative links repectively.

(3)Wikipedia: It is a network derived from the vote information of users of Wikipedia for the election of its administrator. Users can vote for or vote against the candidate which is expressed by positive links or negative links.

(4)Gahuku-Gama: It is a network described the political alliance and hostility of 16 subtribes of the New Guinea highland in 1954. The topology of the network is shown in figure 4. The 16 nodes represent 16 subtribes and 58 edges represent relationships between tribes, of which 29 positive edges shown by solid lines represent alliance and 29 negative edges shown by dashed lines represent hostility.

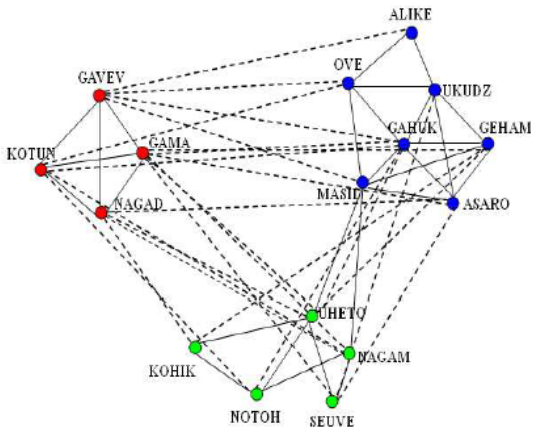

FIGURE 4. Gahuku-Gama subtribes network

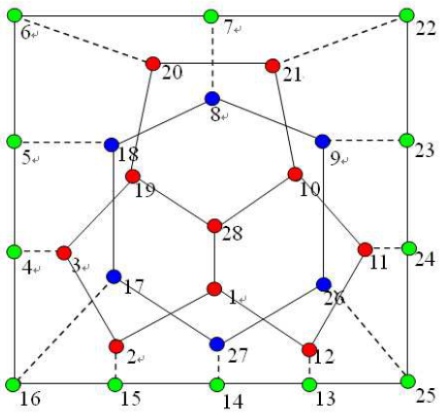

FIGURE 5. The illustrated signed network I

(5) The illustrated signed network I: It is a simulated signed network. Its topology is shown in figure 5 where the solid line represents the positive edge and the dotted line represents the negative edge.

6.2. **Division of training set and testing set.** In order to evaluate the prediction accuracy of the algorithm, we need to divide the known edge set E into a training set and a testing set. In our experiments, we use the ten fold cross method[2] which is commonly used. In terms of each dataset, we randomly choose 10% edges from E as the testing set and denote it by $|E^{Te}|$. The remaining 90% edges are used as the training set which is denoted by $|E^{Tr}|$. Here we should ensure that $E^{Te} \cup E^{Tr}$=E and $E^{Te} \cap E^{Tr} = \varnothing$. The edges in the training set are considered as known information while the testing set is used to test and verify the prediction accuracy of the algorithm. Repeat this division 10 times so as to guarantee that each subset can be used as a testing set only once and all edges in E can not only be trained but also be tested.

6.3. **Choice and improvement of the evaluation index.** There are three commonly used indices for accuracy evaluation of link prediction algorithms, namely AUC, Precision and Ranking Score[2]. The similarity value in the calculation of these indices are positive numbers and these indices can only evaluate the accuracy of link prediction algorithm for traditional social networks or sign prediction for signed networks. However, the link prediction score in PSN_BS can be positive or negative. So, AUC^{BS} and $Precision^{BS}$ were obtained through the improvement of AUC and Precision respectively and were taken as evaluation indices of PSN_BS in accuracy of link and sign prediction.

(1)AUC^{BS} evaluation index: It is used for the evaluation of link prediction accuracy of future edges.

Link prediction score in PSN_BS can be positive or negative. Its absolute value measures the probability of the establishment of the future link and its sign represents the sign prediction result of the future link. Here, we use U represents the edge collection of the complete graph of G and use E^{un}=U-E represents the set that consists of actually nonexistent edges of G. Then in our experiments, in terms of the two edges randomly selected from E^{Te} and E^{un} respectively, the calculation of AUC should be under the condition that the corresponding sign of prediction score of these two node pairs are the same. Orelse, we should terminate this operation and reselect edges to start the next calculation. In the end, AUC^{BS} is defined as shown in formula (6).

$$AUC^{BS} = \frac{n' + 0.5n''}{n} \tag{6}$$

In comparison with AUC described in reference [2], the meaning of parameters in the formula (6) is accordingly adjusted aiming to the characteristic of the sign network. Here, n represents the number of edges that are tested in experiments and it is set to 20000. Each time, we randomly select two edges from E^{Te} and E^{un} respectively, and calculate the prediction score of these two node pairs corresponding to the two edges selected according to PSN_BS algorithm. Only when both of the two scores are positive or negative can we compare their absolute values to calculate AUC^{BS}. Here, if the former is greater than the latter, we would plus 1 to n'. If their absolute values are the same, we would plus 1 to n''. Otherwise, if the sign of two scores are different, we would terminate this operation, reselect edges and start the new calculation.

(2)$Precision^{BS}$ evaluation index: It is used for the evaluation of sign prediction accuracy of existing edges when signs of these edges are unknown.

Precision[2] is commonly used for the evaluation of sign prediction algorithm with time series. While PSN_BS algorithm in this paper is aimed at the static snapshot of the network at some point. So we improved Precision index so as to verify the sign prediction accuracy of the algorithm proposed. In terms of the snapshot of the network graph, we randomly select an edge from the graph each time, take it as the edge waiting for testing, and assume that the edge is nonexistent. Then we predict the sign of the edge according

to PSN_BS algorithm. If the prediction result is the same as the real sign type, it means the prediction is right. Otherwise, the prediction is wrong. In the end, $Precision^{BS}$ is defined as shown in formula (7).

$$Precision^{BS} = \frac{N_{s\_correct}}{N_{s\_total}} \tag{7}$$

Here, N_{s_total} represents the total number of links that were tested and $N_{s\_correct}$ represents the number of links that were correctly predicted in their signs.

6.4. **Evaluation of link prediction accuracy.** In the experiment, different step length connectivity factor is set to verify the effectiveness of the algorithm. Related studies have shown that the path with the shorter step length contributes more to the similarity than the longer path. So in order to reduce the number of calculations, we set λ to 0.5, 0.6, 0.7, 0.8, 0.9 and 1 respectively, and the prediction accuracy of PSN_BS algorithm based on AUC^{BS} is shown in table 4, where the value of AUC^{BS} is the average of the 10 times of experiments.

Table 4. Prediction accuracy of PSN_BS based on AUC^{BS}

Dataset	Link prediction accuracy (AUC^{BS})					
	λ=0.5	λ=0.6	λ=0.7	λ=0.8	λ=0.9	λ=1
(1) Epinions	0.945	**0.965**	0.910	0.922	0.930	0.873
(2) Slashdot	0.912	0.919	0.914	**0.926**	0.909	0.851
(3) Wikipedia	**0.930**	0.929	0.925	0.928	0.919	0.918
(4) Gahuku-Gama	0.738	0.730	0.774	0.812	**0.866**	0.617
(5) Illustrated signed network I	0.5	0.5	0.5	0.5	0.5	0.5

From table 4 we know:

(1) The prediction accuracy of the algorithm can attain 93% or so in terms of the first three datasets. These three networks are real networks commonly used in the research of signed networks and experimental results have shown the ideal performance of the algorithm proposed for link prediction of large scale signed networks. Additionaly, as to the first three datasets, the prediction accuracy reached the maximum value repectively when λ was 0.6, 0.8 and 0.5. It shows that in network of different topological properties, when the proportion of the number of paths connecting the two nodes with step length of 2 and 3 are different, the corresponding balanced triangles and balanced quadruples play different roles in the forming of future links and their signs. So we should choose a reasonable connectivity factor to achieve the best prediction performance in terms of different networks in practical applications.

(2) As to the forth dataset, the overall prediction accuracy is slightly lower compared with the first three datasets. This dataset is very special in its topology as shown in figure 4. The number of positive links and negative links in this network are the same. As to its 16 nodes, there are 3 nodes that the difference between the positive degree and negative degree of each node is 0. There are 5 nodes that the difference between the positive degree and negative degree of each node is -1. There are 2 nodes that the difference between the positive degree and negative degree of each node is -2. There are 2 nodes that the difference between the positive degree and negative degree of each node is -3. And the differences between the positive degree and negative degree of the other 4 nodes are 1, 2, 5 and 7 repectively. The difference between the positive degree and negative degree of all nodes of the network is 0. So in terms of the network with larger average negative density, the algorithm needs to be further improved to increase the prediction accuracy.

(3) As to the fifth dataset, the topology structure is also special as shown in figure 5. The 28 nodes in the network can be divided into two categories according to the distribution of their degrees. Among them, there are 4 nodes that their positive degree is 3 and negative degree is 0. For the other 24 nodes, the positive degree and negative degree are 2 and 1 repectively. So, in calculation of AUC^{BS}, in terms of the two edges randomly selected from E^{Te} and E^{un} each time, the degree distribution and topological propertities of the two corresponding node pairs are almost the same, which leads to the same prediction scores. That is to say, n'$\approx$ 0 and n"$\approx$ n. So AUC^{BS} should also be 0.5 no matter what value of λ is, which further verifies the correctness of the algorithm.

6.5. **Verification of sign prediction accuracy.** In real signed networks, signs of many edges are unknown. For example, the interaction between 80% yeasts in the protein network is unknown. For the network missing sign type, if we can use link prediction algorithm to accurately predict the unknown sign firstly, and then direct the experiment based on these prediction results, it is possible to greatly decrease the number of tests, reduce the cost and accelerate the understanding of the implicit link information in the network. Therefore, it is of great significance to predict these unknown signs.

In order to further verify the effectiveness and correctness of PSN_BS algorithm in the unknown sign prediction, we use $Precision^{BS}$ to evaluate the algorithm in the next experiment. For each dataset, we randomly select an edge from the network each time as the edge waiting for the test, and assume that the edge is nonexistent. Then we calculate $Precision^{BS}$ according to sign prediction results of PSN_BS algorithm where N_{s_total} is set to 20000. Repeat the experiment for 10 times independently to obtain the higher accuracy and the final value of $Precision^{BS}$ is the average of these ten experiments. Here, in order to verify whether the choice of step length connetivity factor is reasonable, we still set λ to 0.5, 0.6, 0.7, 0.8, 0.9 and 1 respectively. The sign prediction accuracy of PSN_BS algorithm based on $Precision^{BS}$ is shown in table 5.

TABLE 5. Sign prediction accuracy of PSN_BS based on $Precision^{BS}$

Dataset	Sign prediction accuracy ($Precision^{BS}$)					
	λ=0.5	λ=0.6	λ=0.7	λ=0.8	λ=0.9	λ=1
(1) Epinions	0.912	**0.936**	0.931	0.917	0.923	0.905
(2) Slashdot	0.834	0.834	0.833	**0.836**	0.830	0.722
(3) Wikipedia	**0.918**	0.910	0.908	0.904	0.899	0.867
(4) Gahuku-Gama	0.833	0.833	0.833	0.833	0.833	0.667
(5) Illustrated signed network I	0.2	0.2	0.2	0.2	0.2	0

From table 5 we know:

(1) The algorithm got higher sign prediction accuracy in the first three datasets and the accuracy reached the maximum value repectively when λ was 0.6, 0.8 and 0.5.

(2) As to the forth dataset, the sign prediction accuracy was all the same when λ was 0.5, 0.6, 0.7, 0.8 and 0.9. Through the analysis we found the total number of paths connecting two nodes in the network with the step length of 2 and 3 is 120 and 132 respectively. That is to say, the contribution of 3-step similarity score in the calculation of the total prediction score is greater than 2-step similarity score. So we adjusted the value of λ in the next experiment and then we found that the accuracy reached the maximum value 1 when λ was 0.4 and reached the minimum value 0.667 when λ was 1. In other cases, $Precision^{BS}$ was always 0.833. The above results further showed that for the network

that the proportion of the number of paths connecting the two nodes with step length of 2 and 3 are different, the value of λ will influence the final prediction accuracy.

(3) As to the fifth dataset, the sign prediction accuracy was poor. Through the analysis we found that in this network the total number of paths connecting two nodes with the step length of 2 and 3 is 112 and 198 respectively. Then we also adjusted λ in the next experiment and found that the accuracy reached the maximum value 0.4 and the minimum value 0 when λ was 0.4 and 1 respectively. In other cases, $Precision^{BS}$ was always 0.2. Moreover, the network is divided into 3 communities, namely, C_1={v_1,v_2,v_3,v_{19},v_{28},v_{12},v_{11},v_{10},v_{21},v_{20} }, C_2={ v_8,v_9,v_{26},v_{27},v_{17},v_{18}}, C_3={ v_4,v_5,v_6,v_7, v_{22},v_{23},v_{24},v_{25},v_{13},v_{14},v_{15},v_{16} }. Nodes in these three communities are represented in red, blue and green color respectively in figure 5. There is no links between C_1 and C_2, but they all connect with C_3 through negative links, and the 12 negative edges are all located between C_1 and C_3 or between C_2 and C_3. So in the experiment, when we randomly select an edge from the network for sign prediction, no matter the positive or negative edge is selected it will always results in that there is no path connecting the tested node pair with step length of 2 and 3, which consequently leads to the prediction score is 0. In this case, we can only predict the sign according to the negative density.

Case I: If the edge selected is any one of the 12 negative edges, the corresponding two nodes will all change into the node that its positive degree is 2 and negative degree is 0. Now the sign prediction result is positive according to the negative density while the real sign of the edge is negative, which means the prediction is wrong.

Case II: If the edge selected is any one of the 30 positive edges, the corresponding two nodes will either all change into the node where their positive and negative degree are all 1 or change into the node where one nodes positive and negative degree are all 1, and the other nodes positive and negative degree are 2 and 0 repectively. In this situation, as for the node pairs that their positive degree and negative are all 1, the negative density of the two nodes is all 0.5 which is larger than the average negative density of the network. Then the sign prediction result is negative which is opposed to the real sign, so the prediction is wrong. As for the node pairs that their positive degree and negative degree are 2 and 0 repectively, the negative density of the two nodes is all 0 which is smaller than the average negative density. Then the sign prediction result is positive that is the same as the real sign, so the prediction is right. While the proportion of these two types of node pairs in the dataset randomly selected is 25/30 and 5/30 respectively, which leads to the final lower sign prediction accuracy of the algorithm. Thus it can be seen that the algorithm needs to be further improved to get more higher sign prediction accuracy in terms of the special signed network where there is no paths connecting the two nodes with step length of 2 and 3.

Additionally, comparing table 4 with table 5 we can see that the change of AUC^{BS} and $Precision^{BS}$ varing with λ are consistent with each other. These two indices all reach the maximum respectively when λ is 0.6, 0.8 and 0.5 in terms of the first three datasets, which further verifies the correctness and effectiveness of the algorithm to combine the similarity with structural balance theory for link and sign prediction in signed networks. Curves of AUC^{BS} and $Precision^{BS}$ varing with λ are shown in figure 6 and figure 7 respectively.

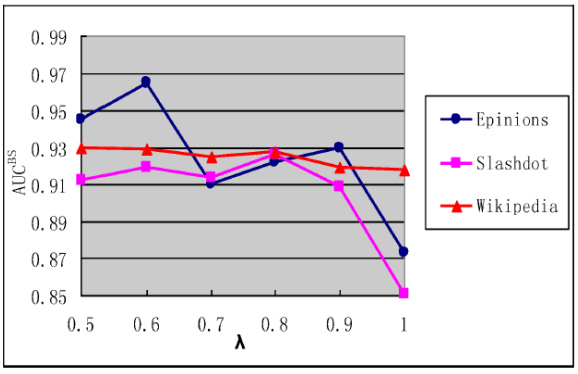

FIGURE 6. Curve of AUC^{BS} varying with λ

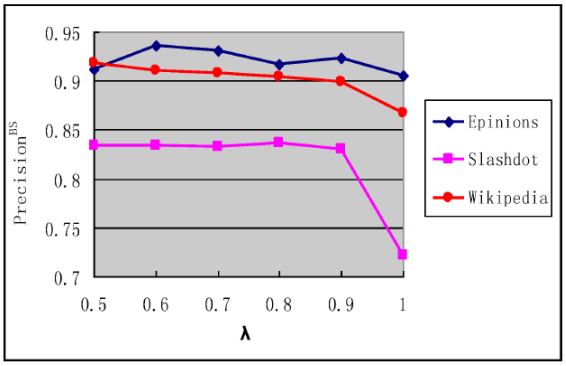

FIGURE 7. Curve of $Precision^{BS}$ varying with λ

6.6. **Comparison with other algorithms.** PSN_BS algorithm can reach the better equilibrium in accuracy and complexity through the combination of local and global similarity of the node on the basis of expanding the length of balanced circles. In order to further verify the performance of the algorithm, we compared PSN_BS with the sign prediction algorithm CN and improved ICN algorithm proposed in reference [21], and used AUC described in that paper as the evaluation index. In the definition of AUC in reference [21], n represents the total number of edges that are tested in the experiment and it is set to |E|, n′ represents the number of positive edges that are correctly predicted and its weight is 1, and n″ represents the number of negative edges that are correctly predicted and its weight is 0.5. Here, according to the above experimantal results, λ was set to 0.6, 0.8, 0.5, 0.4 and 0.4 respectively as for the 5 datasets in the experiment in order to get the higest prediction accuracy. Experimental results of CN, ICN and PSN_BS in prediction accuracy are given in table 6, which have shown that when using AUC[21] as evaluation index, PSN_BS algorithm can always show better performance as to different datasets with a certain degree of stability. And it is superior to CN and ICN algorithm in prediction accuracy.

TABLE 6. Prediction accuracy based on AUC[21]

Dataset	CN	ICN	PSN_BS
(1) Epinions	0.846	0.884	**0.930**
(2) Slashdot	0.759	0.771	**0.914**
(3) Wikipedia	0.868	0.883	**0.938**
(4) Gahuku-Gama	0.732	0.748	**0.866**
(5) Illustrated signed network I	0.500	0.500	**0.500**

7. **Conclusion.** The algorithm PSN_BS based on the similarity and structural balance theory is proposed which take the own attributes of the node and path structure information into account to define the similarity measurement. It improves the deficiency of existing similarity indices on condition of increasing the accuracy and can achieve link and sign prediction simultanously combining with structural balance theory. Experiments have shown the correctness and the higher prediction accuracy of the algorithm proposed using AUC, AUC^{BS}, $Precision^{BS}$ as evaluation indices, which. Moreover, comparison and analysis have also shown that PSN_BS is superior to algorithm CN and ICN. As to large scale signed networks, step length connectivity factor need to be analyzed quantitatively or qualitatively so as to reduce the complexity and improve prediction accuracy of the algorithm proposed.

Acknowledgment. The work was supported by the National Natural Science Fund Project "Research on community discovery and information dissemination mechanism in online social networks based on the theme focused model " (No.61472340), and the National Natural Science Fund Project "Research on the influence of social networks based on information entropy" (No.61602401).

REFERENCES

[1] S. Cheng, H. Shen, G. Zhang and X. Cheng, A research review of signed networks, *Journal of software*, vol. 25, no.1, pp.1-15, 2014.

[2] L. Y. Lv, Link prediction of complex networks, *Journal of University of Electronic Science and Technology of China*, vol.39, no.5, pp.651-661, 2010.

[3] K. Zolfaghar, A. Aghaie, Mining trust and distrust relationships in social web applications, *Proceedings of IEEE International Conference on Intelligent Computer Communication and Processing, New York: IEEE*, pp.73-80, 2010.

[4] M. Lan, C. Li, S. Wang, et al, Research on the prediction algorithm of positive and negative relationships in signed social networks, *Computer research and development*, vol.52, no.2, pp.410-422, 2015.

[5] J. Leskovec, D. Huttenlocher, J. Kleinberg, Signed networks in social media *Proceedings of the 28th International Conference on Human Factors in Computing Systems, New York: ACM*, pp.1361-1370, 2010.

[6] Pranay Anchuri, Malik Magdon-Ismail, Communities and Balance in Signed Networks: A Spectral Approach, *Proceedings of International Conference on Advances in Social Networks Analysis and Mining, New York: ACM*, pp.235-242, 2012.

[7] A. Hassan, A. Abu-Jbara, D. Radev, Extracting signed social networks from text, *Proceedings of the TextGraphs Workshop on Graph-based Methods for Natural Language Processing*, pp.6-14, 2012.

[8] M. Brusco, P. Doreian, A. Mrvar, D. Steinley, Two algorithms for relaxed structural balance partitioning: linking theory, models and data to understand social network phenomena, *Sociological Methods & Research*,vol.40, no.1, pp.57-87, 2011.

[9] M. Malekzadeh, M. A. Fazli, P. J. Khalidabadi, et al, Social balance and signed network formation games, *Proceedings of the 5th Workshop on Social Network Mining and Analysis, NewYork: ACM*, 2011.

[10] M. Szell ,R. Lambiotte, S. Thurner, Multirelational organization of large-scale social networks in an online world, *Proceedings of the National Academy of Science of the United States of America*, vol.107, no.31, pp.13636-13641, 2010.

[11] A. Priyanka, V. K. Garg, R. Narayanam, Link label prediction in signed social networks, *Proceedings of the 23rd International Joint Conference on Artificial Intelligence, California, USA: AAAI Press*, pp.2591-2597, 2013.

[12] C.-J. Hsieh , K. -Y.Chiang, I. S. Dhillon , Low rank modeling of signed networks, *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM*, pp.507-515, 2012.

[13] J. Leskovec, D. Huttenlocher, Jon Kleinberg, Predicting Positive and Negative Links in Online Social Networks, *Proceedings of the International World Wide Web Conference Committee (IW3C2), Carolina, USA: ACM*, pp.641-650, 2010.

[14] K-Y. Chiang , N. Natarajan , A. Tewari , et al, Exploiting longer cycles for link prediction in signed networks, *Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York: ACM*, pp.1157-1162, 2011.

[15] J. Ye, H. Cheng, Z. Zhu, et al, Predicting positive and negative links in signed social networks by transfer learning, *Proceedings of the 22nd International Conference on World Wide Web, New York: ACM*, pp.1477-1488, 2013.

[16] P. Borzymek, M. Sydow , Trust and distrust prediction in social network with combined graphical and review-based attributes, *Agent and Multi-Agent Systems: Technologies and Application*, no.6070, pp.122-131, 2010.

[17] S. Q. Yang, A. J. Smola, B. Long et al, Friend or frenemy?: Prediction signed ties in social networks, *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York: ACM*, pp.555-564, 2012.

[18] G. Facchetti , G. Iacono, C. Altafini, Computing global structural balance in large-scale signed social networks, *PNAS*,vol.108, no.52, pp.20953-20958, 2011.

[19] P. Symeonidis, E. Tiakas, Y. Manolopoulos, Transitive Node Similarity for Link Prediction in Social Networks with Positive and Negative Links, *RecSys*, no.9, pp.26-30, 2010.
[20] A. Patidar, V. Agarwal, K.K. Bharadwaj, Predicting Friends and Foes in Signed Networks using Inductive Inference and Social Balance Theory, *Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp.384-388, 2012.
[21] H. She, M. Hu, Research of link prediction based on signed networks, *Journal of Wuhan University of Technology (Information & Management Engineering)*, vol.37, no.5, pp.464-468, 2015.
[22] E. David, K. Jon, Networks, Crowds, and Markets: Reasoning About a Highly Connected World, *New York: Cambridge University Press*, pp.119-152, 2010.
[23] Z. Weiyu, W. Bin, L. Yang, Integrating Multi-Feature for Link Sign Prediction in Signed Networks, *Journal of Beijing University of Posts and Telecommunications,* vol.37, no.5, pp.80-84, 2014.
[24] M. Liu, J. F. Guo, X. Luo, Link Prediction Based on the Similarity of Transmission Nodes of Multiple Paths in Weighted Social Networks, *Journal of Information Hiding and Multimedia Signal Processing*, vol.7, no.4, pp.771-780, 2016.
[25] Y. D. Li , Community Detection in Signed Networks based on Evolutionary algorithms, *Xi'an: Xi'an Electronic and Science University*, pp.16-18, 2013.