Hindawi
Complexity
Volume 2020, Article ID 4353567, 11 pages
https://doi.org/10.1155/2020/4353567

Research Article

Sign Prediction on Social Networks Based Nodal Features

Xiaoyu Zhu and Yinghong Ma

Business School, Shandong Normal University, Jinan, Shandong, China

Correspondence should be addressed to Yinghong Ma; yinghongma71@163.com

Received 25 April 2019; Revised 12 July 2019; Accepted 25 July 2019; Published 12 January 2020

Academic Editor: Chittaranjan Hens

The sentiments among social individuals are complexity and diversity, and the relationships between them include being friendly and hostile. The positive ("friendly" ,"like" or "trust") or negative ("hostile", "dislike" or "distrust") sentiments in the relations can be modeled as signed connections or links. The missing relations or sentiments between individuals are always worthy of speculation. The sign predication on links has been significant applications in a variety of online settings, such as online recommendation system and abnormal user detections. A novel sign prediction method called the *SPR* model is measured by the values of the two indexes, one is similarity; the other is preference-reputation (PR). The similarity of a pair nodes is defined by the statistical properties of local structures. The definition of similarity agrees with the theory of social balance because existing connections reflect the tendency of the new links emergence between individuals. And PR value is to measure the positive or negative tendency of edges without sign. The experiments on real big social data proved the feasibility and efficiency of the *SPR* model: Comparing with some popular predication methods, the *SPR* model in this issue shows lower complexity and higher accuracy. Experimental results also prove that the *SPR* model provide insight and foresight of the mechanism driving the sign formation of links.

1. Introduction

In social networks, relations among members not only exhibit friendship and cooperation, but also hostility and competition. Positive and negative links were used to describe cooperative (friendly/trustful) and competitive (hostile/distrustful) relationships respectively. Assigning signs to links were a significant way of including additional information to networks than traditional binary or weighted approaches [1–3]. One of the challenges in signed networks is inferring the signs of unknown relations that is often referred to as sign prediction [4], which reveals the underlying relationships between social members. Therefore, it can be widely used in many applications such as recommendation systems and abnormal user detections etc. [5].

Sign prediction is the problem of inferring those hidden signs using the information provided by the rest of the network. It is similar to link prediction, which is a well-studied problem in traditional unsigned social network analysis [6]. However, compared with link prediction, sign prediction is still in its beginning stage due to the following difficulties. One the hand, the effects of negative and positive signs are unbalanced or unwieldy in signed social networks [7, 8]. Positive signs can be propagated between members of social networks while negative signs cannot. For example, A trusts B and B trusts C, A will trust C to some extent, while A distrusts B and B distrusts C, it is hard to judge the relationships between A and C directly [9]. Thereby, in the propagation model of reference [10], the distrust relationship only propagates once among the trust relationships. On the other hand, the formation mechanism of the negative links is different from the positive links. In the field of signed network research, less negative signs datasets are available for study [11] because members of social networks rarely express their antipathy to others for fear of being retaliated [12]. So the negative sign prediction became a difficult problem in the field of sign prediction. Therefore, in-depth study and mining of the formation mechanism of social network is the key to improve the accuracy of prediction.

Sign prediction was first introduced and investigated by Guha et al. [10], and later developed in matrix calculation, machine learning, and collaborative filtering. Guha et al. [10] used power matrix to calculate the propagation of trust and distrust. By the matrix, a variety of technical on predications were discussed. The leading eigenvectors with fitness functions

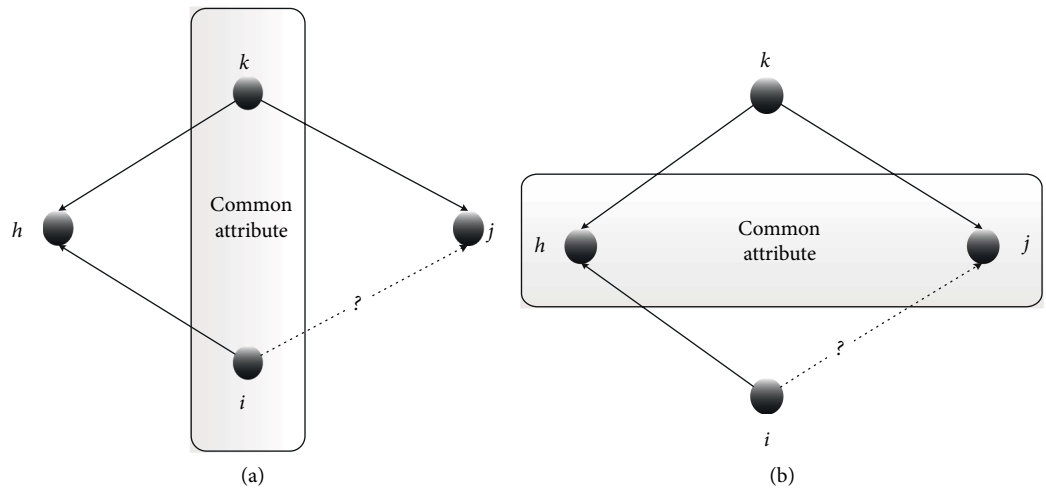

Figure 1: Similarity diagram. (a) The out node pair. (b) The in node pair.

to fine-tune clusters were presented [13]. The random walk according to the similarity between nodal pairs realized in researching the inconsistency of distrust in propagation [14]. Minimizing the rank of the adjacent matrix could approximately make the balanced structure to the greatest extent [15]. To quickly obtain the maximal balanced matrix, Cai et al. [16] propose a singular value projection algorithm, in which the product of the top-k singular vectors and singular values is taken to approximately replace the original matrix. Agrawal et al. [17] and Hsieh et al. [18] approximate the original matrix by a matrix decomposition method, in which the original $n \times n$ matrix is decomposed into the product of two $k \times n$ matrices, and the element values of the product matrix are used as the predicted values. To date, the methods used in machine learning include logistic regression [4, 9, 19, 20], support vector machine [21], decision tree [22], naive Bayes [23] etc.; the features used for learning include nodal degrees [4, 9], types [23], similarity [9, 20], trustworthiness [24], preference [25, 26], triangle structures [4], quadrilateral structures [19], user reviews [22, 27] etc. Collaborative filtering focuses on similarity, similar individuals are more likely to make similar behaviors, which is the basic idea of sign prediction by collaborative filtering. Javari and Jalili [28] believe that computing the similarity between nodes is affected by the sparsity of the social networks. Therefore, they cluster the network and calculate the similarity between clusters to replace the similarity between individuals. Individual behaviors in signed network was believed hidden in "group intelligence" which is embodied by the community structure [5]. The community structure embedded in the social network is untractable even in complete networks [29].

Enlightened by the references and their methods, a new sign prediction method is presented by two indexes in this paper, one is similarity; the other is the preference-reputation (PR) value, called *SPR* model for short. The statistics of local structures are analyzed to explore the constitution mechanism of signed social networks by which the similarity of a pair nodes are defined. The meaning of similarity agrees with the theory of social balance, because the existing connections reflect the tendency of new links emerging between individuals. And the PR value, coinciding with the preferential attachment mechanism [2], is to measure the positive or negative tendency of edges without sign. The experiments on real data proved the feasibility and efficiency of the model. Compared with the popular predication methods, the *SPR* model in this issue shows lower complexity and higher accuracy. Experimental results also prove that the *SPR* model provide insight and foresight of the mechanism driving the sign formation of links.

The arrangement of this paper is follows. The introduction and motivation is illustrated in Section 1; In Section 2, the similarity, and the PR value are defined. Thereafter, the predictive method, namely the *SPR* model, is presented based on the indexes. In Section 3, the experimental results and comparisons on three real social signed networks, Epinions, Slashdot, and Wikipedia, are shown. Finally, the discussion and conclusion of this work are presented in Section 4.

2. The Method and Model

A signed graph is denoted by $G = (V, E, S)$, where V and E are the node set and the link set of G respectively, and $S = \{+1, -1, 0\}$ is a weight set on E such that the link (i, j) is set $s_{ij} = 1$, -1 or 0 if the node i shows positive, negative, or none attitude to the node j. Irrespective of positive or negative, the sentiments are clear and distinct. While, for the none attitude, it is ambiguous and unsetting, people wonder to determine the precise attitude. Then a natural question is to predict the sign of link (i, j) based on the information of E and their signs [4]. The sign prediction problem is also interpreted to "what extent the evolution of a network can be predicted using its structural information" [26].

In this section, indexes such as similarity, dissimilarity, preference and reputation are presented, and the sign of link predication model is constructed.

2.1. Similarity and Dissimilarity. In order to predict the edge sign from node i to node j, s_{ij}, it is necessary to make targeted analysis on the prediction task. Consider the following local structure, as shown in Figure 1: in panel (a), since k is the node into j and $s_{kj} \neq 0$, then the higher common attribute between

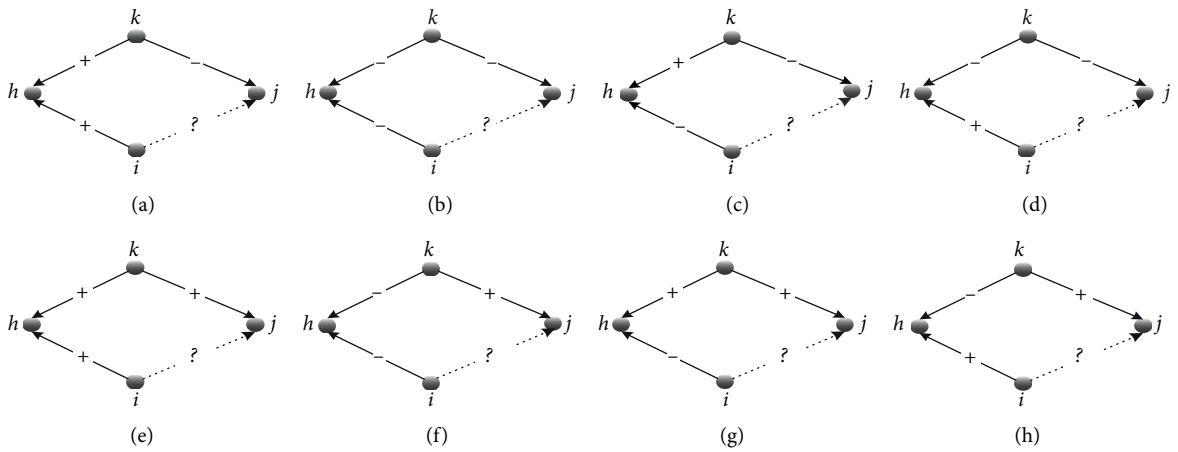

Figure 2: Positive and negative similarity. Panels (a)–(d) denote positive similarity, whereas panels (e)–(h) denote negative similarity.

i and k, the more probability of $s_{ij} = s_{kj}$; in panel (b), since h is the node out of i, then the higher common attribute between h and j, the more probability of $s_{ij} = s_{ih}$. There by predicting s_{ij} can via the common attributes between i and k and the common attributes between h and j. Analyzing Figure 1, since i, k are the source nodes and h, j are the target nodes in the quadrilateral structure, the common attributes between i and k are equal to the common attributes between h and j. Thus, it can yield twice the results with half the effort. Generally, the more common neighbors (polarity is also consistent) two nodes have, the higher their common attributes will be. Then the similarity between i and j can be defined as

$$Sim(i,j) = \frac{\sum_{k\in D_i(j)}\left(\left|D_o^+(i)\cap D_o^+(k)\right| + \left|D_o^-(i)\cap D_o^-(k)\right|\right)}{\sum_{k\in D_i(j)}\left|D_o(i)\cap D_o(k)\right|}, \tag{1}$$

where $D_o^+(i)$ and $D_o^-(i)$ are the neighborhoods getting out the node i with positive and negative links, respectively, $D_i(j)$ is the neighborhoods getting in the node j irrespective of the signs of links. Further, $Sim(i,j)$ is refined by the signs of the node j and its neighbors. Then

$$Sim(i,j) = Sim^+(i,j) + Sim^-(i,j), \tag{2}$$

where $Sim^+(i,j)$ and $Sim^-(i,j)$ are the cases of $k \in D_i^+(j)$ and $k \in D_i^-(j)$ for Equation (1) respectively, and $D_i^+(j)$ and $D_i^-(j)$ are the neighborhoods getting in the node j with positive and negative links respectively. $Sim^+(i,j)$ and $Sim^-(i,j)$ are called the positive similarity and negative similarity, respectively.

Figure 2 shows all the cases of $\sum_{k\in D_i(j)}|D_o(i)\cap D_o(k)|$: where panels (a)–(d) are the case of $k \in D_i^+(j)$ and panels (e)–(h) are the negative similarity $k \in D_i^-(j)$; Hence, panels (a)–(d) show positive similarity $Sim^+(i,j)$, whereas panels (e)–(h) describe the negative similarity $Sim^-(i,j)$. By Equation (1), panels (a) and (b) confirm with $Sim^+(i,j)$, while panels (c) and (d) against it; Panels (e) and (f) confirm to $Sim^-(i,j)$ while (g), (h) are against it respectively. For the opposite property of the similarity, the dissimilarity is also introduced.

In Figure 2, the more structures of (a) and (b), the larger the value of $Sim^+(i,j)$, and the more structures of (c) and (d), the smaller the value of $Sim^+(i,j)$. The more structures of (e) and (f), the larger value of $Sim^-(i,j)$, and the more structures of (g) and (h), the smaller value of $Sim^-(i,j)$.

As the definition of similarity of nodes i and j, the dissimilarity between nodes i and j is defined

$$DSim(i,j) = \frac{\sum_{k\in D_i(j)}\left(\left|D_o^+(i)\cap D_o^-(k)\right| + \left|D_o^-(i)\cap D_o^+(k)\right|\right)}{\sum_{k\in D_i(j)}\left|D_o(i)\cap D_o(k)\right|}, \tag{3}$$

$$DSim(i,j) = DSim^+(i,j) + DSim^-(i,j), \tag{4}$$

where $DSim^+(i,j)$ and $DSim^-(i,j)$ are the cases of $k \in D_i^+(j)$ and $k \in D_i^-(j)$ for Equation (3) respectively, $DSim^+(i,j)$ and $DSim^-(i,j)$ are positive dissimilarity and negative dissimilarity, respectively.

By Equations (1)–(4), it is found that the following two facts hold if $D_o(i)\cap D_o(k) \neq \emptyset$,

$$\begin{aligned} Sim^+(i,j) + DSim^+(i,j) &= 1, \quad \text{if } k \in D_i^+(j), \\ Sim^-(i,j) + DSim^-(i,j) &= 1, \quad \text{if } k \in D_i^-(j); \end{aligned} \tag{5}$$

otherwise, when $D_o(i)\cap D_o(k) \neq \emptyset$, the other two facts hold,

$$\begin{aligned} Sim^+(i,j) = DSim^+(i,j) &= 0, \quad \text{if } k \in D_i^+(j), \\ Sim^-(i,j) = DSim^-(i,j) &= 0, \quad \text{if } k \in D_i^-(j). \end{aligned} \tag{6}$$

Normally, $Sim(i,j)$ represents the degree of consistency between nodes i and j, while $DSim(i,j)$ is the degree of inconsistency between nodes i and j. In real social networks, positive similar nodes tend to have positive relationships, while nodes with large differences between them may have negative relationships.

2.2. Preference and Reputation. In social networks, the preference and reputation of individuals are influential in decision-making to form a connection [25]. The preference, known as optimism or bias in previous studies [26], is for edge generating nodes. Some nodes might be more optimistic than others, meaning their attitude are more likely to be positive. The preference of node i is defined as

$$Pr(i) = \frac{\left|D_o^+(i)\right|}{\left|D_o^+(i)\right| + \left(D_o^-(i)\right)}. \tag{7}$$

$Pr(i)$ measures the general attitude of node i toward other nodes in Equation (7), and also means the probability of positive edges among all edges generated by the node i. The greater $Pr(i)$ is, the higher the probability of node i regenerating another positive edge is.

Reputation, also known as prestige or deserve in previous studies [26], is for edge receiving nodes. Reputation reflects the popularity of a node in the network. A node with a high reputation tends to receive more positive edges. The reputation of node i is defined as

$$Re(i) = \frac{\left|D_i^+(i)\right|}{\left|D_i^+(i)\right| + \left|D_i^-(i)\right|}. \tag{8}$$

In Equation (8), $Re(i)$ measures the general attitude of other nodes toward node i, and it is also the probability of positive edges among all edges received by node i. The greater $Re(i)$ is, the higher the probability of node i receiving another positive edge is.

Combing both $Pr(i)$ and $Re(j)$ would enhance the prediction effect on the pair of nodes i and j. Therefore, we calculate the weighted sum of $Pr(i)$ and $Re(j)$ as

$$PR(i, j) = \frac{\left|D_o^+(i)\right| + \left|D_i^+(j)\right|}{\left|D_o(i)\right| + \left|D_i(j)\right|}. \tag{9}$$

The sum of the coefficients of $Pr(i)$ and $Re(j)$ in Equation (9) is 1, which means the equation not only takes into full consideration the preference of node i and the reputation of node j, but also the priority connection mechanism [2].

2.3. The Prediction: SPR-Model. This section predicts signs using similarity-dissimilarity (denotes as $Sim - DSim$) and PR value. $Sim - DSim$ is a local environmental feature which reflects the interaction structure the target edge actually participated, while PR value is the nodal own feature which reflects the empirical estimates according to the past performances. Here, the prediction method takes both $Sim - DSim$ as the decisive factor and PR value as the auxiliary factor.

The SPR model is taken as follows:

Denote $\Delta^+ = Sim^+(i, j) + DSim^-(i, j)$ as the positive index and $\Delta^- = Sim^-(i, j) + DSim^+(i, j)$ as the negative index. Let ϵ be any given positive real number to measure the difference between Δ^+ and Δ^-, $\epsilon \in [0, 1]$ a threshold measuring the difference between Δ^+ and Δ^-, and ϵ. Δ^+ reflects the positive tendency between nodes, while Δ^- is the negative tendency between nodes. When the gap between Δ^+ and Δ^- is large enough, the tendency is looked as obvious. Therefore, two cases of $\Delta^+ - \Delta^- > \epsilon$ and $\Delta^+ - \Delta^- < \epsilon$ are assumed as the positive and negative signs, respectively. Hence, the sign of the link of nodes i and j is assigned by the two cases:

Case 1. If $|\Delta^+ - \Delta^-| > \epsilon$. In this case, the sign tendency on s_{ij} is easy to understand, so the values of $Sim - DSim$ is competent for the prediction. Therefore, the sign of the link s_{ij} is assigned as

$$s_{ij} = \begin{cases} 1, & \text{if } \Delta^+ > \Delta^- + \epsilon; \\ -1, & \text{if } \Delta^- > \Delta^+ + \epsilon. \end{cases} \tag{10}$$

When $\left|\Delta^+ - \Delta^-\right| > \epsilon$, the sign tendency of s_{ij} is obvious so that the feature $Sim - DSim$ is competent for the prediction task. Yet,

Case 2. $\left|\Delta^+ - \Delta^-\right| \le \epsilon$. This case means that the sentiment's tendency is ambiguous. Hence, the feature of $Sim - DSim$ loses its efficacy for predictions. In this case, the values of $PR(i, j)$ is considered for prediction. Denote the proportion of positive links in the network by p^+. Then the sign of the link s_{ij} is assigned as

$$s_{ij} = \begin{cases} 1, & \text{if } PR(i, j) \ge p^+; \\ -1, & \text{otherwise.} \end{cases} \tag{11}$$

In fact, $PR(i, j) \ge p^+$ means a probability of the preference and the reputation is greater than the proportion of positive tendency, so $s_{ij} = 1$ is easy to admit. Otherwise, $s_{ij} = -1$. When $PR(i, j) = 1$ means the links generated by nodes i and j are all positive; otherwise, the links generated by nodes i and j received are all negative when $PR(i, j) = 0$.

2.4. The Pseudo-Code for Computing the SPR-Model. The pseudo-code for calculating the SPR-model is shown in Table 1.

The computational complexity including time and spatial complexity of the SPR-model algorithm in Table 1 are analyzed. Step 1 computes the nodal neighbor's set by traversing all edges once time, the computational time complexity is $O(|E|)$, where $|E|$ is the size of edge set E; In Step 2, for each edge (i, j), match the neighbors h and k of i and j respectively, the time complexity of Step 2 is $O\left(|E|\langle d\rangle^2\right)$, where $\langle d\rangle$ is the average degree of nodes. In Step 3, computing the similarity and dissimilarity of each pair of nodes takes $O(|E|)$. In Step 4, it takes $O(|E|)$ for computing PR value of each pair of nodes. And finally in Step 5, it also takes $O(|E|)$ for predicting the sign of each edge. Therefore, the total computational time complexity of predicting the signs of edge in E is $O\left(|E|\langle d\rangle^2\right)$.

In the experimental analysis, the input real social networked data is the adjacent matrix with $|E|$ rows times 3 columns. Each row is an edge, the first and the second columns are the source and the target nodes, respectively, the third column is the observed sign from a source to a target node. When we calculate the SPR-model, a $21 \times |E|$ dimensions matrix is defined. As described above, the first three columns are still network link data. The 4th column to the 11th are the number of eight special quadrangles of each edge contained in respectively. The 12th to 15th column store the values of Sim^+, Sim^-, $DSim^+$ and $DSim^-$ of the edge respectively. The 16th to 18th columns are the values of Pr, Re and PR of each edge respectively. The 19th and 20th columns are the values of Δ^+ and Δ^- of each edge respectively. The 21st column is the predicted value for each edge. Hence, the spatial complexity is $O(|E| \times 21)$. In addition, the spatial complexity of calculating the neighbor set of each node is $O(|V|\langle d\rangle)$, where $|V|$ is the

Table 1: Algorithm of pseudo-code of *SPR*-model.

Input: Network adjacent matrix.
Initialization: For each node i, $D_o^+(i) = \emptyset$, $D_o^-(i) = \emptyset$, $D_i^+(i) = \emptyset$, $D_i^-(i) = \emptyset$.
For each edge (i, j), do the following 5 steps:
Step 1. Compute the neighbor set of each node.
If $s_{ij} = 1$, $D_o^+(i) = D_o^+(i) \cup \{j\}$, $D_i^+(j) = D_i^+(j) \cup \{i\}$;
If $s_{ij} = -1$, $D_o^-(i) = D_o^-(i) \cup \{j\}$, $D_i^-(j) = D_i^-(j) \cup \{i\}$.
Step 2. Compute the number of special quadrilaterals.
For $h \in D_o^+(i) - \{j\}$ %The existing sign of s_{ij} is not considered.
For $k \in D_i^+(j) - \{i\}$ %The existing sign of s_{ij} is not considered.
If $h \in D_o^+(k)$, set FIG2(a) = FIG2(a)+1, where FIG2(a) is the number of quadrangles in Figure 2(a).
Similarly, compute FIG2(b), FIG2(c), $\cdots$, FIG2(h).
Step 3. Compute the similarity and dissimilarity.
$Sim^+(i, j) = \frac{\text{FIG2(a)} + \text{FIG2(b)}}{\text{FIG2(a)} + \text{FIG2(b)} + \text{FIG2(c)} + \text{FIG2(d)}}$.
Similarly, compute $Sim^-(i, j)$, $DSim^+(i, j)$ and $DSim^-(i, j)$.
Step 4. Compute PR values.
$PR(i, j) = \frac{\lvert D_o^+(i)\rvert + \lvert D_i^+(j)\rvert}{\lvert D_o(i)\rvert + \lvert D_i(j)\rvert}$.
Step 5. The sign of each edge (i, j) is predicted.
Set $\Delta^+ = Sim^+(i, j) + DSim^-(i, j)$, $\Delta^- = Sim^-(i, j) + DSim^+(i, j)$.
If $\Delta^+ - \Delta^- > \epsilon$, $s_{ij} = 1$;
If $\Delta^- - \Delta^+ > \epsilon$, $s_{ij} = -1$.
If $\lvert\Delta^- - \Delta^+\rvert \le \epsilon$ and $PR(i, j) \ge p^+$, set $s_{ij} = 1$;
If $\lvert\Delta^- - \Delta^+\rvert \le \epsilon$ and $PR(i, j) < p^+$, set $s_{ij} = -1$, where p^+ is the proportion of positive links in the network.
Output: The sign of each edge (i, j).

size of the node set V of the network. Summarizing the above analysis, the total spatial complexity is $O(2|E| + \langle d\rangle|V|)$.

3. Experiments

In order to verify the efficiency and reasonability of the sign of link predication model, experiments on real data are taken. Experiments are included for three real social signed networks, Epinions, Slashdot, and Wikipedia [4]. Epinions is a consumer review site. Users can read or comment on a variety of goods and services, and they can also rate them. Users also can be allowed to evaluate the comments made by other users, that is, evaluate other users as trustworthy or distrusted objects. Epinions dataset consists of 131828 nodes and 841372 edges, 86.0% of which are positive edges. Slashdot is a blog site that allows users to say they like or dislike other users' comments. Slashdot data consists of 82144 nodes and 549202 edges, 77.4% of which are positive edges. Wikipedia is an online voting network where users can vote for or against a candidate administrator. Wikipedia dataset consists of 7118 nodes and 104359 edges, 78.4% of which are positive edges. The details of these three networks are shown in Table 2.

Table 2: Three real social signed networks.

	Epinions	Slashdot	Wikipedia
Nodes	131828	82144	7118
Edges	841372	549202	104359
+Edges (%)	86.0	77.4	78.4
Edges in triangles (%)	80.1	52.1	91.9
Edges extracted (%)	91.6	91.1	97.7

+Edges denote the positive edges in networks.

Table 3: The parameters used to calculate metrics.

		Real	
		Positive	Negative
Predicted	Positive	TP (True positive)	FP (False positive)
	Negative	FN (False negative)	TN (True negative)

Table 4: Three extracted subdatasets.

	Nodes	+Edges	−Edges	% +Edge
Sub-Epinions	82877	657608	113143	85.3%
Sub-Slashdot	54747	383788	116346	76.7%
Sub-Wikipedia	4837	79987	21952	78.5%

+Edges (-Edges) is the number of positive (negative) edges in networks. % +Edge is the percentage of +Edge.

3.1. Evaluating Metrics. Experimental results are presented by three metrics: accuracy, average accuracy and F_1-score. The accuracy (acc) is defined as:

$$acc = \frac{TP + TN}{TP + FP + TN + FN} = \frac{TPR + \rho TNR}{1 + \rho}. \quad (12)$$

where TP, TN, FP and FN are defined as shown in Table 3. TPR is the true positive rate, TNR is the true negative rate, P is the number of positive edges, and N is the number of negative edges. Equation (12) shows that the role of negative edge prediction is almost ignored and the result is completely determined by positive edge when $\rho \to 0$ ($\rho = N/P$). Therefore, the average accuracy ($\overline{acc}$) is defined as:

$$\overline{acc} = \frac{TPR + TNR}{2} = \frac{N \times TP + P \times TN}{2(P \times N)}. \quad (13)$$

Thus, predictors with higher $\overline{acc}$ can predict higher rates of either sign in even skewed datasets disregarding bias [30]. In addition, since sign prediction is a binary classification task, F_1-score is used to measure the predictive precision and recall rate and it is calculated as:

$$\frac{1}{F_1} = \frac{1}{2} * \left(\frac{1}{Precision} + \frac{1}{Recall}\right), \quad (14)$$

where $Precision = TP/TP + FP$ and $Recall = TP/TP + FN$. Obviously, the F_1-score is the harmonic mean of *Precision* and *Recall* and can be a trade-off between them.

3.2. Generalization across Datasets. To test the performance of the predictive model, experiments are made on different

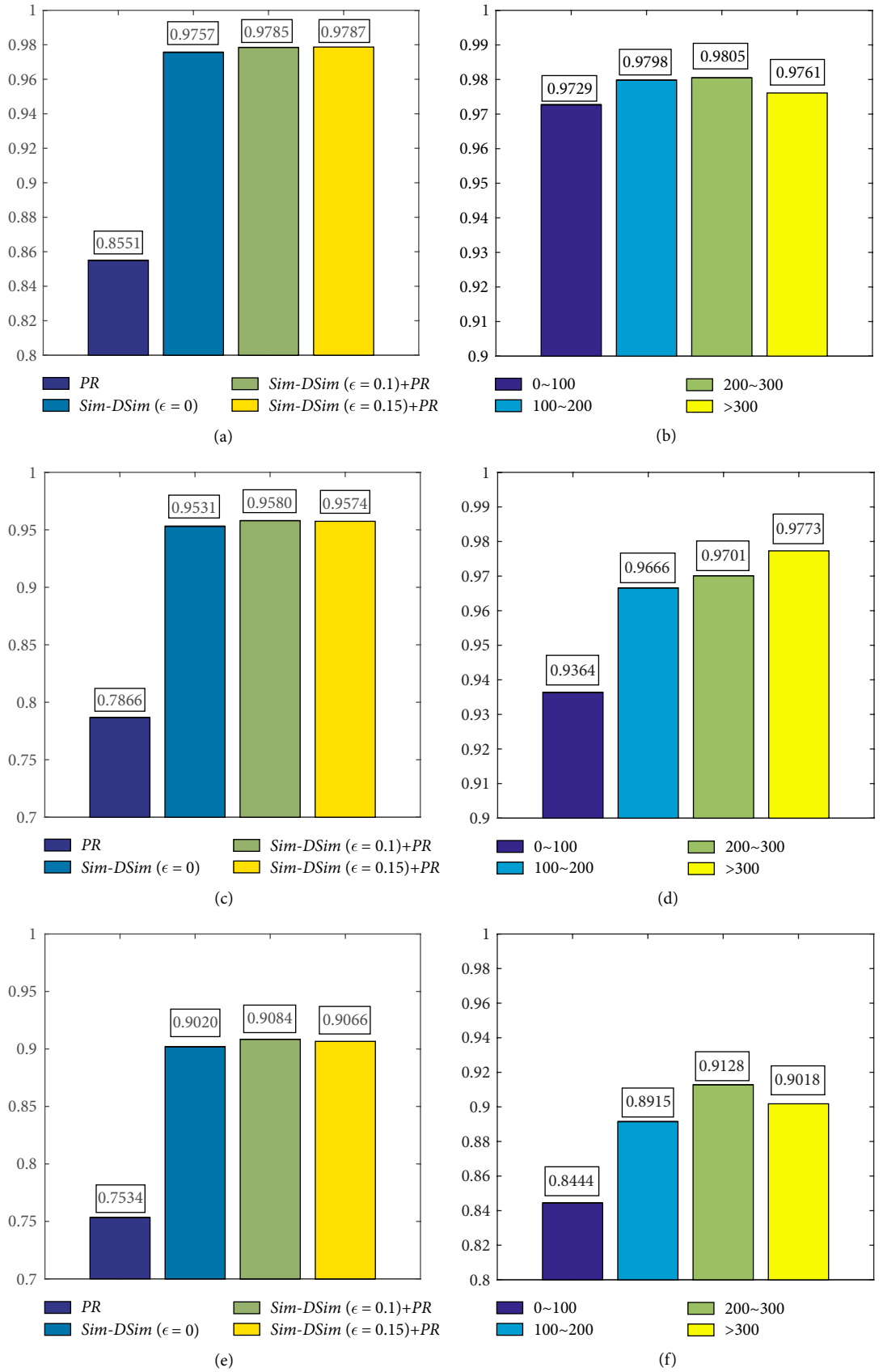

FIGURE 3: Accuracies comparison with different prediction mechanisms or sets' size on the three networks. Panels (a), (c) and (e) are accuracies comparison with different prediction mechanisms; and panels (b), (d) and (f) are accuracies comparison with different set size of each network. (a) and (b) Epinions. (c) and (d) Slashdot. (e) and (f) Wikipedia.

datasets Epinions, Slashdot and Wikipedia. In Table 2, 91.6% of Epinions, 91.1% of Slashdot and 97.7% of Wikipedia are extracted for testing. Table 4 shows the three sub-datasets whose edges are contained in at least one panel of Figure 2.

The performances of the predictive model is displayed in Figures 3(a), 3(c) and 3(e) which demonstrates that: (1) when predicting only based on *PP* value, accuracies on three datasets are 85.51%, 78.66% and 75.34%, respectively, while when predicting only based on $Sim - DSim$, results are 97.57%, 95.31% and 90.20%, improved by 12.06%, 16.65%, and 14.86% respectively. (2) when using $Sim - DSim$ as decisive and *PR* value as auxiliary to predict, accuracies on the three datasets are all improved, which demonstrate the scientific of the predictive model.

Since $Sim - DSim$ is computed by the number of quadrilaterals as Figure 2 displayed, each dataset is classified into four sub-datasets according to the number of quadrilaterals to test the performance of $Sim - DSim$, As shown in Figures 3(b), 3(d) and 3(f). For Epinions, the predictive effect does not differ significantly over the four sub-datasets, moreover, the predictive accuracy always be high. This proves that $Sim - DSim$ has high robustness. For Slashdot and Wikipedia, when the number of quadrilateral is $0:100$, the predictive accuracy is obviously lower than that when the number of quadrilateral exceeds 100. This demonstrates that these two networks have less data to extract features, which is the main reason why the accuracy under these two datasets is not as well as the data of Epinions. Therefore, the conclusions are threefold. First, the network of Epinions is more mature than that of Slashdot and Wikipedia. Second, that the predictive accuracy of Slashdot and Wikipedia increasing with the increased available network data; And the third is scientific to predict with $Sim - DSim$.

3.3. Comparison of Results. To further test the performance of prediction of *SPR* model, it is compared with the existing approaches, such as the logistic regression (LR) proposed by Leskovec et al. [4], the logistic regression based on three attributes (LR-3A) proposed by Yuan et al. [9], the supervised learning based on higher order cycles (HOC) proposed by Chiang et al. [19], the logistic regression based on Bayesian node properties (LR-BNP) proposed by Song et al. [23], the troll-trust model based on ranking proposed by Wu et al. [24], the logistic regression based on reputation and optimism (LR-RO) proposed by Shahriari et al. [26], the measures of imbalance (MOI) and the matrix factorization (MF) studied by Chiang et al. [15], the collaborative filtering (CF) introduced by Javari and Jalili [28] and the closed triple micro structure (CTMS) proposed by Khodadadi and Jalili [30]. The comparison results are shown in Table 5. In order to compare the approaches fairly, the experimental data of Table 5 are quoted from the previous studies. Note that in the predictive model $\epsilon = 0.15$.

Table 5 shows that the *acc* values of SPR-model on Epinions, Slashdot and Wikipedia are all larger than that of other 10 approaches. This proves the feasibility and validity of *SPR*'s predicting mechanism for calculating the nodal features. By comparing the *acc* of the 10 approaches, the following conclusions can be drawn: (1) Social balance theory cannot fully explain the mechanism of the formation of signed social networks, although MOI-10 measures the balance of cycles with lengths ≤ 10, its prediction results are still inferior to those of other algorithms. In addition, the low *acc* of CF also illustrates that the prediction of edge signs should take full account of other features of the network, rather than relying solely on structural balance. (2) Local structure is more signed than macro structure. In other words, nodes generate the signed edges usually based on their local connections, i.e., HOC-5 learns the features of cycles with length of $3:5$, its predictive results are still inferior to those of other machine learning algorithms. (3) Machine learning can not effectively capture the key signed structural features when there are too many features to learn, i.e., for the nine scalars of the three

Table 5: The results of *acc* on three networks.

	Epinions	Slashdot	Wikipedia
LR [4]	0.9342	0.9351	0.8021
LR-3A [9]	0.9592	0.8892	0.8786
HOC-5 [19]	0.9080	0.8469	0.8605
LR-BNP [23]	0.9313	0.8565	0.8737
Toll-Trust [24]	≈ 0.96	≈ 0.90	≈ 0.89
LR-RO [26]	0.9582	0.9010	0.8880
MOI-10 [15]	0.8497	0.7850	0.8220
MF [15]	0.9448	0.8835	0.8839
CF [30]	0.9282	0.8258	0.8137
CTMS [30]	0.9570	0.8598	0.8542
SPR	**0.9664**	**0.9446**	**0.9007**

"≈" is the approximation from the reference.

Table 6: The results of $\overline{acc}$.

	Epinions	Slashdot	Wikipedia
LR-RO [26]	0.9441	0.8975	0.8651
CF [28]	≈ 0.88	≈ 0.83	≈ 0.79
LR [30]	0.7589	0.6887	0.6654
MF [30]	0.8856	0.8217	0.7911
CTMS [30]	0.9083	0.8142	0.7202
SPR	**0.9470**	**0.9387**	**0.8667**

"≈" is the approximation from the reference.

Table 7: The values of F_1-score.

	Epinions	Slashdot	Wikipedia
LR-3A [9]	≈ 0.83	≈ 0.71	≈ 0.78
Troll-Trust [24]	≈ 0.97	≈ 0.94	≈ 0.93
SPR	**0.9802**	**0.9636**	**0.9360**

"≈" is the approximation from the reference.

Table 8: Statistics of bi-directed edges.

	Epinions	Slashdot	Wikipedia
{+1, +1}	97.11%	88.36%	90.54%
{−1, −1}	1.84%	9.60%	3.29%
{+1, −1}	1.05%	2.04%	6.17%

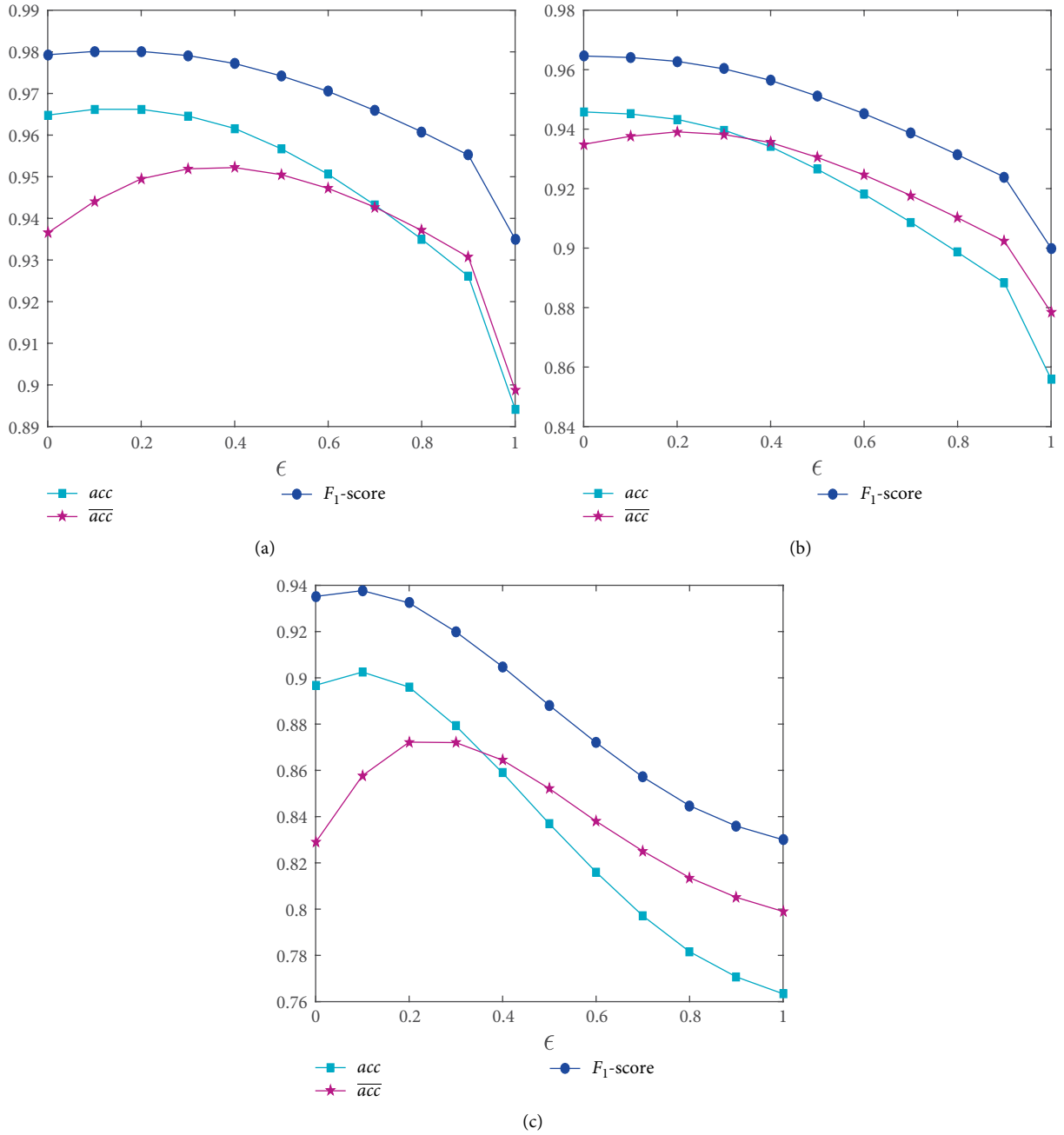

Figure 4: Experimental results of three real data sets under different ϵ. (a) Epinions. (b) Slashdot. (c) Wikipedia.

algorithms (LR, HOC-5 and LR-BNP) there are eight scalars inferior to that of LR-RO. The main reason is that LR-RO only learns two features (reputation and optimism) while the other three algorithms have learnt many features. (4) The main factor affecting the sign of an edge is the features of its two endpoints, followed by its local features, and finally its global features. For these 11 algorithms, there are only Troll-Trust and LR-RO can be comparable to *SPR* in terms of accuracy and robustness. What these three algorithms have in common is that they are based on the features of two endpoints to predict the sign of edge. The above comparative analysis demonstrates that *SPR* successfully avoids the shortcomings of other algorithms and captures the key signed structural features.

As for the skewness feature of actual datasets, *acc* is basically determined by the positive edges. Therefore, the $\overline{acc}$ of the *SPR* model is compared with the exiting algorithms, shown in Table 6. In order to compare the approaches fairly, the experimental data of Table 6 are quoted from previous studies. Since some previous studies did not show the results of these experiments, the kinds of comparison algorithms in Table 6 are less than that in Table 5, and the *SPR*-model significantly outperforms than others showing the scientific and validity of *SPR*'s predictive mechanism. Compared with the five algorithms in Table 6, LR-RO is still the most competitive, which is consistent with the conclusion in Table 5. However, the $\overline{acc}$ of other algorithms has been greatly reduced. This shows that

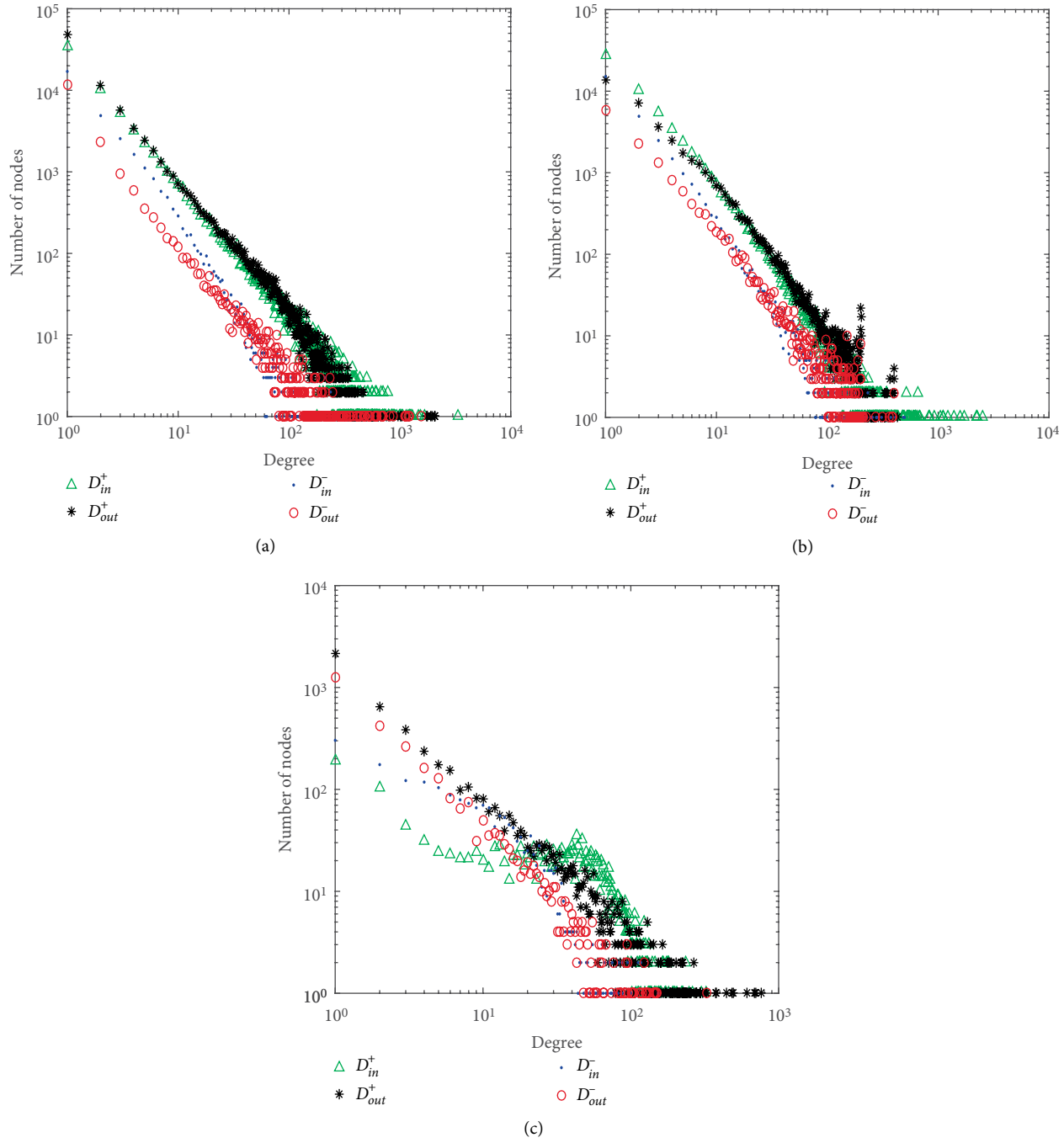

FIGURE 5: Degree distributions of the three real data sets. (a) Epinions. (b) Slashdot. (c) Wikipedia.

most of the algorithms have defects in predicting negative edges. In addition, *SPR*'s F_1-score is also compared with LR-3A and Troll-Trust algorithms, as shown in Table 7, of which the experimental data prove that the predictive model has high predictive precision and recall rate. By comparing with the state of the art methods, it is fully demonstrated that *SPR* outperforms others in predicting both positive and negative edges.

3.4. Analysis of Results. Figure 4 shows the experimental results, plotted as a function of ϵ. With the change in ϵ, the trend of *acc* and F_1-score is basically synchronized, which also shows that the two evaluation metrics are mainly determined by the positive edges, moreover, when ϵ is very small ($0 \le \epsilon \le 0.1$), they can reach the optimum. However, the trend of $\overline{acc}$ is quite different. With the change of ϵ, $\overline{acc}$ shows a clear trend of increasing first and then decreasing, and the optimal value is obviously lagging behind that of *acc* or F_1-score. This is because: when ϵ is very small, the edge signs are mainly determined by the $Sim - DSim$ feature; with the increase of ϵ, a considerable part of the edges are determined by the *PR* value, by this token, *PR* value is superior to $Sim - DSim$ in predicting

negative edges. Yet, due to the overwhelming advantage of the positive edges in quantity, all the three evaluation metrics will be reduced when ϵ is too large.

4. Discussion and Conclusion

In this paper, the *SPR* model is proposed to predict the edge signs in large online social networks where interactions can be both positive and negative. The model is easy to understand because of the only two indexes to measure the interactions between nodes and their local environments.

$Sim - DSim(i, j)$ shows similarity and dissimilarity between nodes i and j, which can be refined into positive and negative similarity-dissimilarity. Experimental results on Epinions, Slashdot, and Wikipedia proved the scientific and validity of $Sim - DSim(i, j)$ in predicting edge signs. The main advantages of the index of $Sim - DSim(i, j)$ precisely predicting edge's signs are as follows. The first advantage is the index of $Sim - DSim$ measuring the common attributes of nodal pairs. Hence $Sim - DSim$ is calculated from a highly symmetrical quadrilateral. Since the signs of the bi-directed edges are basically coincident which powerful supported by the evidences in Table 8. The natural conjecture of the directions of the links in the network should be symmetrical. In fact, the proportions of bidirectional links in Epinions, Slashdot and Wikipedia are 30.55%, 17.39%, and 2.04%, respectively. The reason why Wikipedia has a worse prediction effect might be the bi-direction of links. The second advantage is that the values of $Sim - DSim$ keep both social balance and status theory hold on, or at least it skillfully avoids the conflicts between them. For example, in Figure 2(a), the quadrilateral is structurally balanced when s_{ij} is 1, and s_{ij} should be the same as s_{kj} when nodes i and k have similar status.

The third might no the last advantage is that the *SPR* prediction model makes the best possible of the existing data to predict the missing signs of links. Previous methods are mostly based on triangle structure, and there are fewer triangle data in actual data. As shown in Table 2, in Epinions, Slashdot, and Wikipedia there are 11.5%, 39%, and 5.8% fewer triangle data compared with the data the model based.

$PR(i, j)$ displays the tendency of s_{ij}, and is a weighted sum of the preference of i and the reputation of j. Nodal preference and reputation are derived from the preferential attachment mechanism, which can be described in signed social networks as: nodes with larger positive/negative outdegree (or indegree) generate a positive/negative edge with larger probability; nodes with smaller positive/negative outdegree (or indegree) generate a positive/negative edge with smaller probability, shown in Figure 5. Experimental results demonstrate that negative edges have obvious *PR* features when they are generated. Therefore, it may be more effective to predict edge signs by distinguishing the features of nodal pairs.

In this paper, the underlying mechanism that determine the signs of links in large social networks is explored and a conclusion is obtained that edge signs are mainly determined by their own or local features, not the global one. Through experimental analysis, the scientificity and validity of the predictive model are verified. In addition, because the features measured by the model are extracted from the nodal own or local structures, the model is very advantageous for large-scale datasets.

Data Availability

The three .txt files, Epinions.txt, Slashdot.txt, and Wikipedia.txt are datasets used to support the findings of this study have been deposited in the Stanford web site repository at https://snap.stanford.edu/data/#signnets. The datasets are in the form of adjacency list, include three arrays: the first is the source node, the second is the target node, and the third is the edge weights or the signs. The data of Epinions is the consumers' review site, includes 131828 nodes and 841372 links. Users can read or comment on a variety of goods and services, and also be allowed to evaluate the comments made by others users, that is, to evaluate other users as trustworthy or distrusted objects. The data of Slashdot is a blog site that allows users to say what did they like or dislike other users' comments, and it contains 82144 nodes and 549202 links. The data of Wikipedia is an online voting network where users vote or against a candidate administrator, and there is 7118 nodes and 104359 links.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We would like to thank the anonymous reviewers for the constructive comments and suggestions, which undoubtedly improved the presentation of this paper. We show our great appreciation to all the authors who collected and shared the data, such as Epinions, Slashdot and Wikipedia to be benchmark networks. Finally, we would like to thank the National Science Foundation of China (No. 71471106) that supported this research.

References

[1] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[2] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[3] E. Z. Erbach-schoenberg, S. Bullock, and S. C. Brails-ford, "A model of spatially constrained social network dynamics," *Social Science Computer Review*, vol. 32, no. 3, pp. 373–392, 2014.

[4] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th International Conference on World Wide Web–WWW '10*, pp. 641–650, 2010.

[5] G. K. Awal and K. K. Bharadwaj, "Leveraging collective intelligence for behavioral prediction in signed social networks through evolutionary approach," *Information Systems Frontiers*, vol. 21, no. 2, pp. 417–439, 2017.

[6] D. Liben-Nowell and J. M. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 1019–1031, 2007.

[7] G. Costa and R. Ortale, "Model-based collaborative personalized recommendation on signed social rating networks," *ACM Transactions on Internet Technology*, vol. 16, no. 3, pp. 1–21, 2016.

[8] H. Deng, P. Abell, O. Engel, J. Wu, and Y. Tan, "The influence of structural balance and homophily/heterophobia on the adjustment of random complete signed networks," *Social Networks*, vol. 44, pp. 190–201, 2016.

[9] W. Yuan, C. Li, G. Han, D. Guan, L. Zhou, and K. He, "Negative sign prediction for signed social networks," *Future Generation Computer Systems*, vol. 93, pp. 962–970, 2019.

[10] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust," in *Proceedings of the 13th International Conference on World Wide Web ACM*, pp. 403–412, 2004.

[11] S. Wang, J. Tang, C. C. Aggarwal, Y. Chang, and H. Liu, "Signed network embedding in social media," in *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 327–335, 2017.

[12] W. Yuan, D. Guan, Y. Lee, S. Lee, and S. J. Hur, "Improved trust-aware recommender system using small-worldness of trust networks," *Knowledge-Based Systems*, vol. 23, no. 3, pp. 232–238, 2010.

[13] Y. Ma, X. Zhu, and Q. Yu, "Clusters detection based leading eigenvector in signed networks," *Physica A: Statistical Mechanics and its Applications*, vol. 523, pp. 1263–1275, 2019.

[14] M. Jamali and M. Ester, "TrustWalker," in *Proceedings of the 15th ACM: SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 397–406, 2009.

[15] K. Chiang, C. Hsieh, N. Natarajan, I. S. Dhillon, and A. Tewari, "Prediction and clustering in signed networks: a local to global perspective," *Journal of Machine Learning Research*, vol. 15, pp. 1177–1213, 2014.

[16] J. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[17] P. Agrawal, V. K. Garg, and R. Narayanam, "Link label prediction in signed social networks," in *International Joint Conference on Artificial Intelligence*, pp. 2591–2597, 2013.

[18] C. Hsieh, K. Chiang, and I. S. Dhillon, "Low rank modeling of signed networks," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '12*, pp. 507–515, 2012.

[19] K. 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 – CIKM '11*, pp. 1157–1162, 2011.

[20] 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.

[21] A. Nguyenthi, P. Q. Nguyen, T. D. Ngo, and T. Nguyen-hoang, "Transfer adaboost SVM for link prediction in newly signed social networks using explicit and PNR features," *Procedia Computer Science*, vol. 60, pp. 332–341, 2015.

[22] K. Zolfaghar and A. Aghaie, "Mining trust and distrust relationships in social web applications," in *Proceedings of the 2010 IEEE 6th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 73–80, IEEE, New York, NY, USA, 2010.

[23] D. Song and D. A. Meyer, "Link sign prediction and ranking in signed directed social networks," *Social Network Analysis and Mining*, vol. 5, no. 1, p. 52, 2015.

[24] Z. Wu, C. C. Aggarwal, and J. Sun, "The troll-trust model for ranking in signed networks," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining – WSDM '16*, pp. 447–456, 2016.

[25] M. Shahriari and M. Jalili, "Ranking nodes in signed social networks," *Social Network Analysis and Mining*, vol. 4, no. 1, p. 172, 2014.

[26] M. Shahriari, O. A. Sichani, J. Gharibshah, and M. Jalili, "Sign prediction in social networks based on users reputation and optimism," *Social Network Analysis and Mining*, vol. 6, no. 1, p. 91, 2016.

[27] S. Yang, A. J. Smola, B. Long, H. Zha, and Y. Chang, "Friend or frenemy?" in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '12*, pp. 555–564, 2012.

[28] A. Javari and M. Jalili, "Accurate and Novel Recommendations," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 4, pp. 1–20, 2014.

[29] Y. Ma and X. Zhang, "Estimating the number of weak balance structures in signed networks," *Communications in Nonlinear Science and Numerical Simulation*, vol. 62, pp. 250–263, 2018.

[30] 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.