CHARACTERIZATION AND EDGE SIGN PREDICTION IN SIGNED NETWORKS

Tongda Zhang, Haomiao Jiang, Zhouxiao Bao

{tdzhang, hjiang36, zhouxiao}@stanford.edu

ABSTRACT

In this project, we perform an intensive study on online social networks in which the relationships between entities can be either positive which indicates relations such as trust or friendship or negative which represents relationships such as opposition or antagonism. We investigate some basic characteristics for signed networks, make an extension on prior work, propose creative features, and modify the PageRank algorithm to make it applicable in signed networks. We also build up an edge sign prediction model using supervised machine learning results. The experimental results show our model can significantly improve the prediction accuracy and decrease the false positive rate.

***Index Terms*—** Signed Network, Prediction of Edge Sign, Local Bias, SN-PageRank, Machine Learning

1. INTRODUCTION

1.1 Background

In recent years, social network has become an increasingly important resource to analyze individuals' behaviors and the embedded communal structures. In most networks, edges simply indicate connections between nodes. However, sometimes more information about the connection needs to be represented and thus signed network was proposed.

In signed network, each edge is either positive or negative depending on whether it indicates positive or negative information. Positive information mainly includes trust, likes or approves while negative sign generally represents distrust, dislikes or denounce. Signed networks are generally used to characterize attitudes among a group of people. Actually, unsigned networks can be viewed as a special case of signed network, a network with only positive edges. Hence, the algorithms for signed networks can be applied to general networks. Typical signed network includes Epinions, Slashdot and Wikipedia.

With rich information contained in signed network, we can not only make macro analysis on the evolution and structure of real-world social network but also uncover the hidden relationship between two given nodes. Indeed, later research can be categorized into these two directions: global structure modeling and signed edge prediction. For macro analysis, balance and status theories have been proposed in [2] and further studied in [3, 8]. For edge prediction, Guha et. al proposed an algorithm based on exponentiating the adjacency matrix in [6] and Jure et. al took it one step further by using machine learning scheme. Besides optimizing total accuracy, some papers, like [5], put their focus on improving the false positive rate.

1.2 Our Work

The purpose of our project is to conduct an intensive analysis on signed networks. We explore several basic descriptors for signed networks and their variants. Among them, a modified version of PageRank – SN-PageRank – is proposed, which can be successfully applied to signed networks. We also verify our findings to real-world datasets.

We first explore a set of characteristics for signed networks. In addition to some basic features commonly used in network related algorithms such as clustering coefficient, we also extend the triangle patterns adopted by Jure in [3]. We intend to catch an intuitive idea about their relative importance and representativeness of corresponding node and the whole network. After complete calculation in real-world datasets, we build up a collection of features containing various amount of information of the network.

From our knowledge, traditional PageRank cannot be directly applied to signed networks because of the negative edges. Their existence makes it impossible to guarantee the convergence of PageRank algorithm. We modify the original algorithm according to the characteristics of signed network itself, and then apply it to the real-world datasets. Based on the specific modification, we analyze from an intuitive respective, connecting the SN-PageRank with the reliability of every node.

The proliferation of online signed networks and social network analysis lead researchers to the problem of predicting the sign of an edge in signed networks [3, 4, 5]. We follow this line to implement sign prediction based on our discoveries. In this problem, we are following an experimental framework articulated by Jure et al. in their study [3].We successfully propose an edge sign prediction model which integrates the information of local bias and the modified PageRank values for each node. By performing a series of experiments, we find that our model significantly outperforms previous work in prediction.

Positive and negative signs have different meanings in different datasets. For example, the sign of an edge represents trust or distrust of one user toward another user in

Epinions; while in Slashdot, a network of technology blogs, the sign means approval or disapproval of one user of another user's comments. We use one dataset to train our model and test on the other one. We notice that the model is dataset independent without significant deterioration.

As noted above, there are a number of features mentioned in the system. It is valuable the find out which of them are significant and which of them are negligible when predicting edge signs. In order to know this, we implement forward feature selection on the whole feature set. We are capable to clearly distinguish between them after running the algorithm.

1.3 Paper Structure

The remainder of this paper proceeds as follows. Section 2 gives a brief description of a list of prior work which is related to our project. Section 3 focuses on analyzing the dataset we use - Epinions and Slashdot, from their statistical information and some basic network descriptors. Section 4 describes three newly proposed descriptors of signed networks and elaborates on their practical meaning in signed networks. In section 5, we refine the supervised machine learning model for edge sign prediction and perform some experiments on different datasets. We also implement feature selection and cross dataset validation of the model. In the last part, we draw some conclusions based on the simulation results.

2. RELATED WORK

In section 1.1, we briefly introduced some main directions and most remarkable work. Here, we will give more details and survey further lines of study that are also related. Also, we will explain what's the relationship and difference between their ideas and our work.

First, one of our goals is to characterize the signed social network with a meaningful model and use it infer the sign of a given edge. Some previous paper also did this and use prediction accuracy as their goals. Guha et. al introduced belief propagation concept in [6] and used exponential of adjacent matrix as features of their model. They enumerate some possible values for their parameters and get the optimal prediction accuracy for their model. Years after, in [3], Jure et. al made some extensions to it and proposed a model based on logistic regression scheme. In that paper, local information, such as in-and-out degrees, is added. We are going to follow their steps whereas we will predict with brand new features under various machine learning scheme.

Second, we take reliability of each node into consideration. All of previous papers about signed network thought all nodes and their connections are reliable. However, in rating network, like Epinions, the edge sign is highly affected by their personal interests as well as their appreciation abilities. So, we need to characterize the user reliability. This idea is similar to PageRank, a well-defined algorithm used in web searching applications. PageRank algorithm is introduced and fully developed in paper [13, 14, 15] and it can iteratively compute the authority and reliability of each node. However, the original PageRank algorithm can only be applied to directed graph. We'll build up our PageRank for signed network and use it as one important feature in our prediction model.

Finally, we care about generalization problem. To avoid overfitting and make the model easy to analyze, several methods have been proposed. Cross validation is the most commonly used one and is illustrated in [16]. Jure et. al generalized this idea to cross-set validation in [3], which can be used to find the common structure in online network. Besides cross-validation, feature selection is also widely used. Forward feature selection proposed in [11, 12] is an aggressive algorithm to find the smallest feature set with a relatively high accuracy. In this paper, we'll combine these two methods to refine our prediction model.

3. DATASET ANALYSIS

We consider two large online social networks where each link is explicitly labeled as positive or negative to perform our analysis and experiments: Epinions and Slashdot.

3.1 Dataset Introduction

Epinions is a website where users write reviews about a variety of topics. Users can write new reviews, rate the reviews of others, and indicate trust or distrust to another user. Trust information is used to determine the rank of items in certain category and also the rank of reviews for one specific item. With trust information, most valuable or convincing reviews are presented to the user. Reviewers at Epinions are paid royalties based on how many times their reviews are read, which can encourage user to write more reviews. In order to avoid junk reviews and improve the quality, concept of distrust was introduced. With the help of trust or distrust, the web is able to select and present most useful reviews to the user.

Slashdot is a technology-related news website owned by the Dice Holdings, Inc.. Summaries of stories and links to news articles are submitted by Slashdot's own readers, and each story becomes the topic of a threaded discussion among users. Discussion is moderated by a user-based moderation system. Randomly selected moderators are assigned points which can be used to rate a comment. Moderation applies either −1 or +1 to the current rating, based on whether the comment is perceived as either normal, off topic, insightful, redundant, interesting, or troll.

Our experiments and analysis will be performed on these two datasets. Their statistics and characteristics of the data will be described in section the rest parts of section 3.

3.2 Basic Analysis

Some basic statistics of dataset Epinions and Slashdot are shown in Table 1.

Table 1: Statics of Epinions and Slashdot

Property	Epinions	Slashdot
# Nodes	131828	82140
# Edges	841372	549202
# Pos Edges	717667(85%)	42507 (77%)
# Neg Edges	123705(15%)	12410 (22%)
Largest SCC	41441	27382
Largest WCC	119130	82140
Cluster Coef	0.128	0.059
Transitivity	0.081	0.024

3.3 Degree Distribution

Degree distribution of the signed network is also an interesting factor to consider. The log-log plots for degree distribution of the two networks are shown as below:
As can be seen, the degree distribution of signed network obeys the power law distribution. After fitting the data, we found that Xmin = 2 and $\alpha = 1.72$ for Epinions network. For Slashdot data, Xmin = 1 and $\alpha = 1.49$.

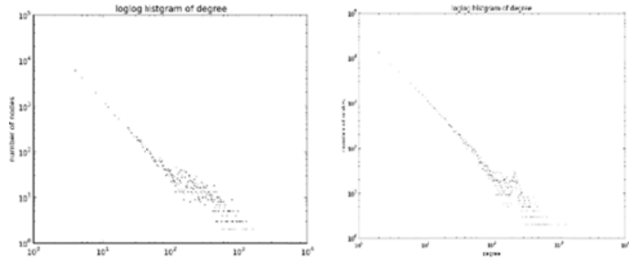


Fig.1 Degree Distribution of Epinions and Slashdot

Another interesting property about signed network is balancing. Balancing is a special case of our lossy belief / trust propagation model in the scenario of three nodes (namely triangles). Here, we only care about eight types of triangles with directed edges and among them, 88% of total triangles are balanced and others are not. So, balancing theory is supported by the data of Epinions.

We first need to count the number of loops of different patterns. In the table, the label of each line represents the signs of edges in triangle ABC. p stands for positive and n stands for negative. *Loop* and *Prop* determines the edge direction. *Loop* triangle means $(A \to B, B \to C, C \to A)$ while *Prop* means $(A \to B, B \to C, C \to A)$. We only list the table for Epinions.

Table 2: Number of Triangle Patters in Epinions

	Loop	Prop		Loop	Prop
ppp	6,096,310	9,594,233	nnp	69,050	52,385
pnp	259,837	328,616	pnn	69,050	271,530
npp	259,837	207,201	npn	69,050	296,078
ppn	259,837	129,852	nnn	16,757	82,094

From the table above, we can see that the success rate of belief / trust propagation is very high (around 88.78%). This shows that trust or distrust can propagate through routine / loops of length 3.

For routines of length four, namely rectangles, the counting information for all the patterns are listed below. The labels of each line take the same rule described above.

Table 3: Number of Quadrangle Patters in Epinions

	Loop	Prop		Loop	Prop
pppp	195,481,458	407,547,491	pnnp	4,114,491	17,975,843
nppp	15,869,360	40,589,460	pnpn	3,170,792	1,664,807
pnpp	19,077,360	24,501,936	ppnn	3,744,978	14,915,703
ppnp	15,869,360	9,952,785	nnnp	672,188	2,137,360
pppn	19,077,360	16,926,874	nnpn	898,121	586,094
nnpp	3,744,978	7,552,592	pnnn	898,121	6,731,991
npnp	2,040,988	2,606,922	npnn	672,188	2,381,472
nppn	4,114,491	4,433,569	nnnn	341,496	2,181,778

From the table above, we can see that trust / distrust can still somehow transmit through routine of length 4. The average success rate is around 76%.

4. CHARACTERIZE SIGNED NETWORK

4.1 Local Bias

Consider an edge from node a to node b in a signed network, which can be represented as (a, b). We are going to take the meaning behind the sign into account here. Actually, the sign reflects the trustiness of a for b in Epinions and likeness of a for b in Slashdot. We will only use Epinions to illustrate below.

Intuitively, for one node, its in-coming positive and negative connections denote the trustiness and distrust it receives from others. The out-coming positive and negative edges represent the trustiness and distrust that node expresses to others. Therefore, we have the following definitions for any node in a signed network:

Incoming Local Bias (ILB): the percentage of negative reviews it receives in all the incoming reviews;
Out-coming Local Bias (OLB): the percentage of negative reviews it gives in all of its out-coming reviews.

$$\mathrm{ILB}(\mathrm{a}) = \mathrm{P}\{\mathrm{sign}(\mathrm{b} \to \mathrm{a}) = -, (\mathrm{b}, \mathrm{a}) \epsilon \mathrm{E}(\mathrm{G})\}$$
$$\mathrm{OLB}(\mathrm{a}) = \mathrm{P}\{\mathrm{sign}(\mathrm{a} \to \mathrm{b}) = -, (\mathrm{a}, \mathrm{b}) \epsilon \mathrm{E}(\mathrm{G})\}$$

For simplicity of expression, we define four values:

$$\mathrm{IP}(a) = \sum\nolimits_{b \in \{(\mathrm{b},\mathrm{a}) \epsilon \mathrm{E}(\mathrm{G})\}} 1\{\mathrm{sign}(\mathrm{b} \to \mathrm{a}) = +\}$$
$$\mathrm{IN}(a) = \sum\nolimits_{b \in \{(\mathrm{b},\mathrm{a}) \epsilon \mathrm{E}(\mathrm{G})\}} 1\{\mathrm{sign}(\mathrm{b} \to \mathrm{a}) = -\}$$

$$\mathrm{OP}(a) = \sum_{b \in \{(\mathrm{a,b}) \epsilon \mathrm{E(G)}\}} 1\{\mathrm{sign(a \rightarrow b) = +}\}$$

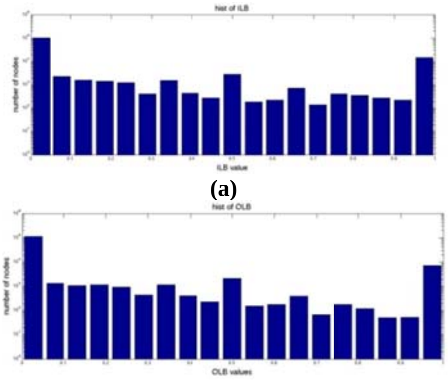
$$\mathrm{ON}(a) = \sum_{b \in \{(\mathrm{a,b}) \epsilon \mathrm{E(G)}\}} 1\{\mathrm{sign(a \rightarrow b) = -}\}$$

Using maximum likelihood estimation, we can get:

$$\mathrm{ILB(a)} \approx \frac{\mathrm{IN}(a)}{\mathrm{IN}(a) + \mathrm{IP}(a)}$$

$$\mathrm{OLB(a)} \approx \frac{\mathrm{ON}(a)}{\mathrm{ON}(a) + \mathrm{OP}(a)}$$

The histograms of ILB and OLB in Epinions are shown in Fig. 2. Slashdot has similar results.



(a)

(b)

Fig.2 Histograms of (a) ILB and (b) OLB

However, when using local bias to predict edge sign, the sign of the objective edge is actually unknown before the prediction. So the previous calculation formula needs to be modified a little to eliminate the information provided by the objective edge.

Two approaches can be adopted to achieve this goal. The first solution is to do the calculation by pretending that the objective edge does not exist. The second solution is to assume that the objective edge can be positive or negative with equal probability, i.e. the edge has sign + with probability 0.5 and has sign – with probability 0.5. Experimental results show that second approach is able to generate better prediction result than the first one since it contains the connection information.

4.2 SN-PageRank

Essentially, when predicting the sign of edge $(\mathrm{a,b})$, the previous proposed characteristics ILB and OLB can be regarded as the attempt to utilize the big number theory to estimate the relative trustiness between a and b from local information of an edge. Take a step further from the section 4.1, a naturally extension is to find out the absolute trustiness using the global network information. The global approach is assuming that each node in the network has its absolute trustiness value that can be compared to others.

From the concept stated in the section 4.1, positive and negative edges transmit trustiness and distrust (a minus trustiness value), respectively. Resembling to background of PageRank algorithm, power iteration method seems to be an applicable approach to get the absolute trustiness value for each node in the signed network. We name it as SN-PageRank (PageRank for signed network).

Similar to traditional PageRank algorithm, we start with definitions, analogy to PageRank. Define the absolute value of trustiness of a node i as t_i, and the stochastic adjacency matrix $\mathrm{M} \in \mathbb{R}^{n \times n}$ where

$$M_{ij} = \begin{cases} \dfrac{\mathrm{sign(i \rightarrow j)}}{\mathrm{out\_degree}(i)} & if\ \mathrm{i \rightarrow j} \\ 0 & otherwise \end{cases}$$

Define the random teleport rate $1 - \beta$, then we have the same iteration algorithm:

$$t^{(k+1)} = \beta M \cdot t^{(k)} + (1 - \beta) \cdot \mathbf{1}$$

where

$$\mathbf{1} = [1\ 1 \ldots 1]^T$$

Unfortunately, the situation is different from the original PageRank since the entries of the stochastic adjacency matrix can be both positive and negative. This will make this algorithm problematic because the sum of a row will not be unitary anymore if an edge with a negative sign shows up.

Here we address this problem step by step:
Assume the original adjacent matrix is $\mathrm{A} \in \mathbb{R}^{n \times n}$ where

$$a_{ij} = \begin{cases} 1 & if\ i \rightarrow j\ and\ sign(i \rightarrow j) = + \\ -1 & if\ i \rightarrow j\ and\ sign(i \rightarrow j) = - \\ 0 & otherwise \end{cases}$$

Actually, the matrix is:

$$\mathrm{A} = \begin{pmatrix} 0 & sign(1{,}2) & \ldots & sign(1, n) \\ sign(2{,}1) & 0 & \ldots & sign(2, n) \\ \vdots & \vdots & \ddots & \vdots \\ sign(n, 1) & sign(n, 2) & \ldots & 0 \end{pmatrix}$$

To remove the negative values in the adjacent matrix without impacting the relative trustiness relation of each pair of nodes, we add 1 to all the entries of A. Therefore, each negative edge is represented by a “0”, no connection is expressed as a “1”, and a positive edge is recorded as a “2”. Hence, every link is added by one, and the relative value transmission is not changed after this modification.

Now we have modified adjacency matrix as:

$$\mathrm{A} = \begin{pmatrix} 1 & sign(1,2)+1 & \dots & sign(1,n)+1 \\ sign(2,1)+1 & 1 & \dots & sign(2,n)+1 \\ \vdots & \vdots & \ddots & \vdots \\ sign(n,1)+1 & sign(n,2)+1 & \dots & 1 \end{pmatrix}$$

To this point, negative edges are removed and the power iteration can theoretically get to work. But the original sparse matrix becomes a dense one. This is caused by the fact existed in real-world datasets that the number of pairs with negative relationship is much less than that of irrelevant pairs. This makes the whole power iteration computationally inefficient / infeasible. So we make a further approximation operation aiming at reducing the dense matrix to sparse matrix without significant impact on the result.

Our approach is cycle cancellation, in which the redundant cyclic paths will be cancelled out. Since after removing the negative edges, a huge number of positive edges are added to the network and a lot of cycles are generated, trustiness will be transmitted back and forth through all those cycles. In this case, the reciprocal relationship is only slightly changed which makes the canceling intuitively a good approximation. At the same time, this method eliminates the extra linking information, making it more consistent with the original graph.

Mathematically, we define the cycle cancellation operation on a matrix M is:

$$\mathcal{K}(\mathrm{M}) = \frac{1}{2}[(M - M^T) \circ (M - M^T) + (M - M^T)]$$

where $\circ$ is the Hadmard product. So the new adjacency matrix we get is $A' = \mathcal{K}(\mathrm{A})$. Now, the matrix is sparse and can be used to the power iteration. So the stochastic adjacent matrix we used in the power iteration is:

$$M = A' \cdot diag\left(\frac{1}{\sum_j a'_{1,j}}, \frac{1}{\sum_j a'_{2,j}}, \dots, \frac{1}{\sum_j a'_{n,j}}\right)$$

The histogram of SN-PageRank (name this result as *SNPR* in following validation part) obtained using this approach for Epinions is shown in Fig. 3.

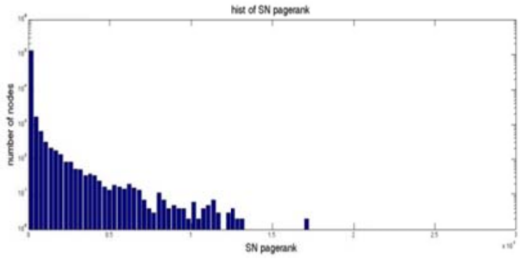


Fig.3 Histogram of SN-PageRank

To validate this approximation approach, we separate the original network into a positive network and a negative network. And we perform the power iteration for both graphs. Let $\mathrm{T} = (t_i)$ be the global trustiness for the positive network and $\mathrm{D} = (d_i)$ be the global distrust for the negative network. We can get a transferred vector as $\mathrm{TD} = \mathrm{T} - \mathrm{D}$. To make TD comparable with *SNPR*, the following region transformation is needed:

$$(\mathrm{TD})_i = \alpha \cdot (\mathrm{TD})_i + bias$$

where

$$\alpha = \frac{[\max(SNPR) - \min(SNPR)]}{[\max(\mathrm{TD}) - \min(\mathrm{TD})]}, bias = \min(SNPR) - \min(\mathrm{TD}).$$

The distribution of TD is shown in Fig.4.

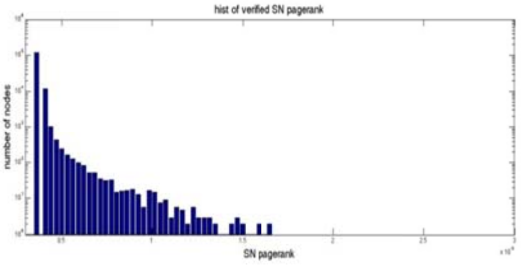


Fig.4 Histogram of TD

From the two plots, we can see that they have fairly similar distributions although certain bias does exist. Therefore, the SN-PageRank is a reasonable estimation of the global trustiness in singed network.

4.3 Weighted Local Bias

In section 4.1, we define the local bias for certain node. By using that concept, we assume that each incoming positive edge is equally important in affecting the local bias of a single node with incoming negative edge. This assumption is indeed the estimation of true bias based on the big number theory. Hence, if the number of neighbors of a given node is not large enough, the experimental results using OLB and ILB may not be satisfactory.

With the help of SN-PageRank values defined in section 4.2, we can recalculate the local bias from a more global aspect.

Instead of directly calculating the percentage of negative incoming edges, we can calculate the corresponding ratio weighted by the SN-PageRank value of that node. The four values defined in section 4.1 are updated as:

$$\mathrm{IP}(a) = \sum_{b \in \{(\mathrm{b,a}) \epsilon \mathrm{E(G)}\}} SN\_PR_b \cdot 1\{\mathrm{sign(b \to a) = +}\}$$

$$\mathrm{IN}(a) = \sum_{b \in \{(\mathrm{b,a}) \epsilon \mathrm{E(G)}\}} SN\_PR_b \cdot 1\{\mathrm{sign(b \to a) = -}\}$$

$$\mathrm{OP}(a) = \sum_{b \in \{(\mathrm{a,b}) \epsilon \mathrm{E(G)}\}} SN\_PR_a \cdot 1\{\mathrm{sign(a \to b) = +}\}$$

$$\mathrm{ON}(a) = \sum_{b \in \{(\mathrm{a,b}) \epsilon \mathrm{E(G)}\}} SN\_PR_a \cdot 1\{\mathrm{sign(a \to b) = -}\}$$

SN_PR_b is just the SN-PageRank value for node b calculated in section 4.2. By applying them to the formula of ILB and OLB, we can get the weighted local bias, which are named as WILB. Since the weights added to $\mathrm{OP}(a)$ and $\mathrm{ON}(a)$ are the same, which cause WOLB remains the same as OLB. Hence, we will not define WOLB redundantly. The distribution of WILB is shown in Fig.5.

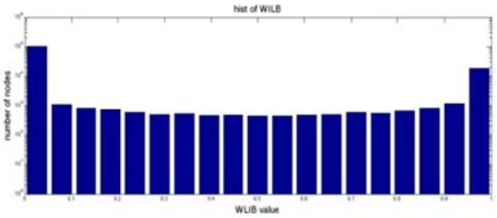


Fig.5 Histogram of WILB

5. EDGE SIGN PREDICTION

5.1. Supervised Method

In this project, we use two different types of supervised classifier to predict the edge sign: logistic regression classifier and SVM classifier.

5.1.1 Logistic Regression

We first use a logistic regression classifier to combine the evidence from useful features into an edge sign prediction. Logistic regression leans a model of the form:

$$\Pr(\mathrm{e}|\mathrm{x}) = \frac{1}{1 + \exp[-(b_0 + \sum_i^n b_i x_i)]}$$

In prediction process, predicted sign of edge e is set to 1 if $\Pr(\mathrm{e}|\mathrm{x}) > 0.5$ and set to 0 otherwise.

5.1.2 SVM Algorithm

We then apply the SVM algorithm to find the hyper-plane with largest by solving the following optimization problem:

$$\min_{\theta} \frac{1}{2} \|w\|^2$$
$$\mathrm{s.t.}\, y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1, i = 1,2, \dots, m$$

After successfully finding out the optimal hyper-plane, we are able to give every objective edge a sign according the relative position of it with respect to the plane.

5.2 Feature Extraction

In the two adopted machine learning algorithms, using what features influences much on the final results. The whole set that we construct mainly contains three kinds of features. First, we consider some commonly used dynamics for network and signed network such as node degree and clustering coefficients. Second, we extend the features adopted by previous work. By combining the idea of exploring longer cycles in [5] and the idea of triangle patters in [2], we explore cycles with length of four. Finally, we propose a modified version of existing algorithm (PageRank), which can be successfully applied to signed networks and has relatively important role in the entire feature set.

In the following parts, we perform a serious of experiments based on this feature set and make a brief comparison between related features. Also, to catch an idea about their relative significance in edge sign prediction, we also implement the feature selection algorithm.

5.3 Various Result

5.3.1 Prediction Accuracy

We use logistic regression and SVM to train and test our model. We follow the experimental scheme in Guha's paper and learn models on original as well as sampled datasets with 50% positive edges. The results under 10 fold cross validation is shown in Table 4 below.

Table 4: Prediction Accuracy using Different Algorithms

Triangle Only				
	Epinions		SlahDot	
Baseline	50%	85%	50%	77%
Logistic	83.46%	91.67%	63.55%	80.84%
SVM	77.47%	89.56%	63.59%	81.06%
Triangle + Quadrangle				
	Epinions		SlashDot	
Baseline	50%	85%	50%	77%
Logistic	87.66%	93.62%	77.28%	83.90%
SVM	85.59%	91.29%	74.94%	82.32%
All				
	Epinions		SlashDot	
Baseline	50%	85%	50%	77%
Logistic	91.38%	95.02%	84.56%	87.13%

As can be seen, the experimental results are encouraging and can support our ideas of quadrangle counting and SN-PageRank estimation. From the table, we can arrive at the points below:

- When dataset size is not large enough, logistic regression generally works better than SVM. SVM cannot get a reasonable hyper-plane when the number of features is small.
- Quadrangle information can greatly improve the prediction accuracy. Although prediction accuracies

when only triangles or quadrangles are almost the same, their information is somehow complimentary.

- With our proposed SN-PageRank, we can push our prediction accuracy to a quite high level. Our model can reach 91.38% on sampled Epinions dataset, which is significantly better than 86.7% from Guha's paper. It is also slightly better than Jure's model on undivided data (i.e., em = 0).

5.3.2 False Positive Rate

Besides precision, recall is also a very important criterion in machine learning problems. In our datasets, positive edges is overwhelm larger than the number of negative edges. As a consequence, the false rate tends to be high. Using our model with all features, the false positive rates for the two datasets are 23.97% and 36.26%, which is much better than 44.4% and 50.7% reported in [5].

5.4 Feature Selection

After training process, we have a well calibrated model, which can be further used to analyze the importance of the features. Observing the weighting coefficients is the most widely used method and it requires carefully normalization on the dataset. However, this requirement cannot be easily fulfilled in our settings. Hence, we used another option, forward feature selection, to analyze the importance of features and simplify the model.

Feature selection process on Epinions dataset is shown in Fig.6 below.

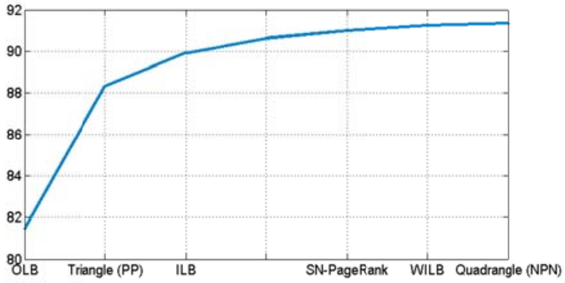


Fig.6 Results of Feature Selection

From the figure, we can see that OLB, ILB and SN-PageRank are important. Also, quadrangle information is also useful in edge sign prediction. Similarly, OLB, ILB and Quadrangle information stands out in process of feature selection on Slashdot dataset.

5.5 Cross Dataset Validation

Since online social network shares similar structures, we can generalize our model and fit it to the whole online social network.

To evaluate this kind of generalization, we train our model on one dataset and test it one the other one. The generalization accuracy for the 50%-positive sampled data is shown in Table 5 below.

Table 5: Generalization Accuracy

	Epinions (Test)	**Slashdot (Test)**
Epinions (Train)	91.3%	82.19%
Slashdot (Train)	88.81%	84.56%

From the table, we can see that our model can be generalized to analyze many online social networks. Since we are using dataset with embededness equals to zero, we cannot make direct comparison with the results in [3].

CONCLUSION

In this project, we investigated several different sorts of characteristics of signed networks, including some commonly used dynamics, extension of triangle patterns, local descriptors and a modified version of PageRank value. By feed some of the selected features to the edge sign prediction model, we successfully improve the performance of previous work to a large extent, in terms of both the prediction accuracy and the false positive rate. By conducting the feature selection algorithm, we find that the local bias, quadrangle and the SN-PageRank can make some contribution in predicting edge signs.

There are a number of further directions can be investigated starting from this project. The first one is to explore other creative and effective methods that might yield still better performance than the edge sign prediction model used in this work. And then, based on current feature set, it is possible to explore more representative features or descriptors that might work well specifically for signed networks. Then, the feature selection algorithm can also be applied to this set to find out the important features and negligible ones. Based on the results, it is fairly valuable to combine them with the social theories of signed links and build up a thorough understanding on why the features are significant. Another important aspect is that, as is mentioned in section 4.1, the meaning of edge signs varies much in different social network; it is worthy to explore our methods and the extensions using other datasets.

AKNOWLEDGEMENTS

We thank Ashton Anderson for all his valuable suggestions and discussions during the process of our project. And also, thanks Jure Leskovec for giving the lectures about social networks, especially the signed networks. Thanks our classmates for information sharing and answering our questions on Piazza.

REFERENCES

[1] D. Cartwright and F. Harary, Structure balance: A generalization of Heider's theory, *Psychological Review*, 63(5):277–293, 1956.

[2] J. Leskovec, D. Huttenlocher, and J. Kleinberg, Signed networks in social media, *In CHI*, pages 1361–1370, 2010.

[3] J. Leskovec, D. Huttenlocher and J. Kleinberg, Predicting positive and negative links in online social networks, *In WWW*, pages 641–650, 2010.

[4] Z. Lu, B. Savas, W. Tang, and I. S. Dhillon, Supervised link prediction using multiple sources, *In ICDM*, pages 923–928, 2010.

[5] K. Chiang, N. Natarajan, A. Tewari, Exploiting longer cycles for link prediction in signed networks, *In CIKM*, October 24-28, 2011.

[6] R. V. Guha, R. Kumar, P. Raghavan and A. Tomkins, Propagation of trust and distrust, *In Proc. 13th WWW,* 2004.

[7] P. Doreian, A. Mrvar, A partitioning approach to structural balance, *Social Networks 18*, 1996.

[8] S. Marvel, S. Strogatz, J. Kleinberg, Energy landscape of social balance, *Physical Review Letters*, 103, 2009.

[9] A. Clauset, C.R. Shalizi, and M.E.J. Newman, Power-law distributions in empirical data, *SIAM Review 51(4)*, 661-703, 2009.

[10] B.A. Huberman, L. A. Adamic, Growth dynamics of the World Wide Web, *Nature 399*, 1999.

[11] F.C. Garcia-Lopez, M. Garcia-Torres, B. Melian, J.A. Moreno-Perez, J.M. Moreno-Vega, Solving feature subset selection problem by a Parallel Scatter Search, *European Journal of Operational Research*, vol. 169, No. 2, pp. 477-489, 2006.

[12] Peng, H.C., Long, F., and Ding, C., Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp. 1226–1238, 2005.

[13] J. Kleinberg, Authoritative sources in a hyperlinked environment, *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[14] A. Altman, M. Tennenholtz, Ranking systems: the PageRank axioms, *Proc. of ACM EC*, 2005.

[15] Z. Gyongyi, H. Garcia-Molina, J. Pedersen, Combating web spam with TrustRank, *Proc. of VLDB*, 2004.

[16] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12)*, 1137–1143, 1995.