# A comparison of methods for link prediction in signed networks

Social Network Analysis for Computer Scientists — Course Project Paper

Zhenyu Guo
LIACS, Leiden University
z.guo.3@umail.leidenuniv.nl

Yanfang Hou
LIACS, Leiden University
y.hou.2@umail.leidenuniv.nl

## ABSTRACT

Our project is to study the problem: *How to infer the relationship between two nodes in signed networks with reliable methods?* To address this problem, we investigate the methods from Leskovec et al. [5] and Chiang et al. [1] and apply these models on the new datasets. We train an appropriate model for link prediction by extracting degree and k-cycle based features from signed network structures. We compare the performance of the models with different feature combinations and the results prove that the longer cycle features will explain the data properties better and produce better prediction results. Also, the difference between our experimental results with that of the original papers will be discussed in depth.

## Keywords

relationship, sign values, logistic regression, features

## 1. INTRODUCTION

With the rapid development of social networks websites, many researchers have studied the features of various types of social networks. The signed network research is an essential subject to investigate the relationships among entities [5]. In contrast with former researches, recent signed network studies concentrate on both negative and positive relationships. For instance, the link signs in Reddit hyperlink networks indicate whether the source subreddit express the positive or negative attitudes towards the target subreddit in the post. In news review websites such as Slashdot, users can express their agreement and disapproval for the news or the articles. The entity scores will become lower if they receive many critical comments and doubts in community. Thus, it is necessary to consider both relationships in social networks. Reliable inferences for the link signs in networks is the goal for Leskovec et al.[5] and Chiang et al [1]. Concretely, their problem is to predict the binary relationships (positive or negative) between two nodes in real-world datasets. This problem is also named as *edge sign prediction problem*.


*SNACS '18* Social Network Analysis for Computer Scientists, Master CS, Leiden University (`liacs.leidenuniv.nl/~takesfw/SNACS`).

Leskovec et al. study the edge sign prediction problem and address the problem via a supervised learning method [5]. They investigate the network features for their machine learning method with two different respects, common features and features based on social psychology study. Common features refer to various *degree* type features between two nodes. They reflect the positive or negative relations of an node to the rest of network [5]. Features based on social psychology study are denoted by triad type features based on the *social balance* theory. More specifically, social balance follows some common principles, such as "friend of my friend is my friend", "enemy of my friend is my enemy" and "enemy of my enemy is my friend". In light of this theory, we apply the various triads as our second type of features. The experiment results indicate their contributions in tackling the edge sign prediction problem via machine learning method.

However, the signed network datasets in the real world are usually sparse and many nodes do not have common neighbors, which makes the triangle-based features not work efficiently. The model performs worse when there are no common neighbors between two nodes. For this issue, Chiang et al. [1] introduce high-order cycle features into the model. Concretely, they add longer cycle based features, such as a quadrilateral, a pentagon and other polygon cycles. In addition, they use the imbalanced dataset instead of proposed balanced dataset to finish the experiment. The fact that their method outperforms Leskovec's methods in each dataset proves that higher-order cycle features could improve the prediction results.

To better understand the link sign prediction problem, we start with following the methods proposed in [5] and [1] to reproduce the experimental results. We will evaluate the two models on the signed social networks with different embeddedness levels, which indicates the minimum number of common neighbors around two nodes within an edge in the edge set. Apart from the three datasets used in [5] and [1], we will apply the supervised model to another two new datasets and observe their performance. Finally, due to imbalance in the number of positive and negative edges in the graph, we follow the paper [1] to add false positive rate as indicator.

The structure of this paper can be divided into several parts.

We will firstly introduce the motivation for this project and then state our problem. And then we will describe the proposed methods for the problem in detail. Next, we will set up the experiment and make a brief introduction for the datasets and metrics. The experiment results will also be illustrated and we will compare the results of two methods by two metrics. At last, we will draw a conclusion for this project and list some points that can be improved in the future.

## 2. RELATED WORK

Many studies address the link sign prediction problem from different perspectives. The two classical socio-psychological theories, social balance and status theory, are commonly appeared in the related studies. From a social balance point of view, Leskovec et al. propose that the sign prediction for a given edge should minimize the number of unbalanced triangles, while for social status, the objective is to assign each node a unique status so that more edges point from low to high [5].

Some researches apply supervised learning method to predict link sign by treating it as a binary classification problem. As mentioned in the introduction part, Leskovec et al. formulates a generalized supervised model by extracting triangle-based features in addition to the degree features from the graph [5]. Chiang. et al. incorporate the high-order cycle into features to further improve the model performance [1]. Zhang et al. incorporates the local bias (the percentage of negative reviews an edge gives or received) and the SN-PageRank (an modification of PageRank) into features and found that they could be helpful in predicting the edge signs [7].

Unsupervised learning methods are also used to solve the link sign prediction problem. Javari et al. propose a new method that first partition the signed network into a set of clusters and then utilize a collaborate-filtering algorithm for sign prediction[4]. In [3], Hsieh et al. models the link prediction as a low-rank matrix completion problem based on weak structural balance in signed networks and successfully recover the missing relationships by leveraging the state-of-art matrix completion algorithm.

## 3. PROBLEM STATEMENT

To formally define the link sign prediction problem, we first consider a directed signed graph $G = (V, E)$ where $V$ denotes the set of nodes and $E$ indicates the set of signed edges between nodes. We use $s(u, v)$ to represent the sign of edge $(u, v)$. $s(u, v) = 1$ if there is a positive edge from $u$ to $v$, $s(u, v) = -1$ if the sign is negative and $s(u, v) = 0$ if there is no connection from $u$ to $v$. The graph $G$ has an adjacency matrix $A \in \{-1, 0, 1\}^{|V|\times|V|}$. We define the positive part $A^+$ and negative part $A^-$ as: $A^+ = max(A, 0)$ and $A^- = min(A, 0)$, where $max$ or $min$ are applied element-wise.

A cycle in a graph is a path of non-zero length from $v$ to $v$ with no repeated edges. A simple cycle is a cycle with no repeated vertices except for the beginning and ending vertex. In this paper, we focus on the simple cycles. The term k-cycle represents a simple cycle consisting of k nodes.

Given a directed network with a sign on each edge, We first divide the edges of into the training and test set. Assuming the edge signs in the training part are given to us, our problem is to infer the edge signs in the test set.

## 4. METHODS

We formulate a supervised model proposed by [5] for link prediction in signed networks. Both of models are based on logistic regression but with different features. The first model construct features from degree and triangles [5] and the other model incorporates the features from long order cycles [1].

### 4.1 Logistic Regression

As our project is to predict the binary values of edge signs in the networks, we apply the logistic regression model for training and evaluation. The logistic regression can be formulated as:

$$P(+|x) = \frac{1}{1 + e^{-(\theta_0 + \sum_{i=1}^{n} \theta_i x_i)}}$$

where the outcome will be interpreted as the probability of a positive link. The $x_i$ and $\theta_i$ values represent the extracted features and coefficients of these features, respectively. The feature extraction will be described below.

### 4.2 Degree and triad based features

It is essential for feature selection because features that we extract from data will form the logistic regression model and influence the prediction results.

We will extract two types of features for our supervised model. The first class of features we extract from the data are based on degrees, which will depict the accumulated relations between nodes and others in the networks. For example, if we want to understand A's evaluation to B, we could get relevant information from A's reviews to others and other reviews received by B. Specifically, for a given edge from $u$ to $v$, we extract the following seven degree-based features.

- $d^+_{in}(v)$ and $d^-_{in}(v)$: the number of incoming positive and negative edges to $v$.
- $d^+_{out}(u)$ and $d^-_{out}(u)$: the number of outgoing positive and negative from $u$.
- $C(u, v)$: the embeddedness of the edge $(u, v)$. $Embeddedness$ denote the common neighbors of $u$ and $v$ regardless of directions.
- $d_{in}(v)$ and $d_{out}(u)$: the total indegree of $v$ and the total outdegree of $u$, which equal $d^+_{in}(v) + d^-_{in}(v)$ and $d^+_{out}(u) + d^-_{out}(u)$ respectively.

The second class of features are based on social balance theory, which follows the common principles, such as "the enemy of my friend is my enemy" and "the friend of my friend is my friend". It means we could predict the relationship between two persons through their relations with the third person. In this work, we utilize the triad types of features to

represent the relationships between the common neighbors and two investigated nodes. After counting both directions and signs for links between common neighbors and the investigated two nodes, 16 types of triad features will be extracted. The more concrete details for triad type feature can be given as: Suppose we should predict the edge sign value between node $A$ and node $B$, and there is a node $C$ which connect to both nodes and the link sign values are known (shown in Figure 1). We can infer the link sign between node $A$ and node $B$ through the known link signs between node $C$ and both of them. Besides, the links can be in either direction and of either sign value. The total amount of triad type features will be 16.

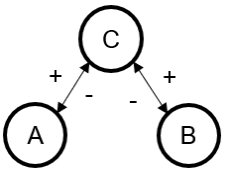



Figure 1: Triad type

## 4.3 Longer cycles based features

The above model does not achieve good performance when many nodes in the graph do not have common neighbors. In this case, the prediction for the edges with zero-embeddedness totally relies on the degree-based features, which will introduces a bias in the learning. Thus the model does not perform robust for the social network in which many people do not share friends. To address this problem, [1] expands the circle size and adds the longer cycle based features. We can also give some intuitions for such a change. For example, if A's friend and B's friend are friend, then A and B are likely to be friend.

The triad-based feature corresponds to k-order cycle for $k = 3$. Considering the sign and direction of each edge in a triad, there is a total of 16 triad types. Similarly, there are 64 four-order cycle types under $k = 4$. The higher order cycles are analogous to these types. The Figure 2 illustrate the higher order cycle types.

The k-order cycle based features could be obtained by the matrix power, which states that the entry $i$, $j$ in $A^k$ counts the number of walks from $i$ to $j$ of length $k$ [2]. According to this theory, for the edge $(i, j)$, the k-order cycle features can be obtained as the $(i, j)$ entries in the $4^{k-1}$ matrices:

$$(A^{b_1})^{t_1} \cdot (A^{b_2})^{t_2} \cdot ... \cdot (A^{b_{k-1}})^{t_{k-1}} \quad (1)$$

where $b_i \in \{\pm\}$ indicates whether we use the positive or negative part of $A$ and $t_i \in \{T, 1\}$ represents whether we transpose the matrix $A$.

# 5. EXPERIMENTS

We conduct the experiments on the five real-world datasets across different fields, which are Epinion, Slashdot, Wiki-Rfa, Bitcoin and Reddit datasets. The linear classifier uses the following features respectively: (1) degree and triad based features; (2) based on the features mentioned in (1), we add the quadrangle based features. In addition, we train the model by using the edge sets with different minimum

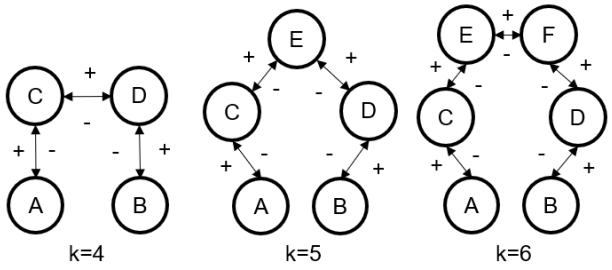



Figure 2: Longer cycles

embeddedness. Finally, we adopt the 10-fold cross validation method to evaluate the model performance in terms of accuracy and false positive rate.

## 5.1 Datasets

The five signed networks datasets will be used for comparing the primary and improved sign prediction methods. The first three datasets are applied in the original papers [5][1]. In addition, the two new datasets are added to investigate the generalization problem of the methods. All the datasets can be found on *snap.stanford.edu* [6].

- *Epinions*: this is a review website for different products with a very active user community. Users will express their thoughts for the comments of the products to indicate whether they trust the product reviews or not. The trust and distrust will represent the positive and negative relationship between users.

- *Slashdot*: this is a technology news website and users in the community can leave their likes or dislikes for other users' comments to indicate whether the users are friends or enemies.

- $wiki - Rfa$: this dataset comprises the links between voters and candidates for community administration status election. The link signs indicate whether the users support or oppose the candidates to become the administration status. It is worth mentioning that we remove the edges with zero sign value because zero sign value denotes the voter's neutral attitude towards the candidate, which is infeasible in this binary prediction task.

- *Bitcoin*: due to the anonymous property of Bitcoin OTC, this dataset consists of the edges between users and the trust levels between two traders so that they can prevent risky traders. As the trust levels range from -10 to 10, we process the ratings to -1 and 1 values. The processing principle is converting all positive weights to 1 and others to -1.

- *Reddit*: a dataset that represent whether the source subreddit is positive or negative towards the target subreddit. The sign values 1 and -1 indicate the positive and negative attitude respectively. With regard to replicated links, we merge the links to the unique and sum the sign values. Then we convert these sum values to +1 and -1 according to the signs of themselves.

We do some statistics about the number of nodes and edges for the above datasets, shown in the table 1. Besides, figure 3 shows the fraction of edges with different levels of minimum embeddedness $E_k$. From $E_k = 0$ to $E_k = 5$, there is a significant drop in the number of edges, indicating the fact that the signed networks in the real word are usually sparse and many nodes do not share neighbors.

Table 1: Dataset statistics

| | Nodes | Edges | + edges | - edges |
|---|---|---|---|---|
| Epinion | 131828 | 841372 | 85.3% | 14.7% |
| Slashdot | 82140 | 549202 | 77.4% | 22.6% |
| Wiki-Rfa | 11255 | 176953 | 78.3% | 21.7% |
| Bitcoin | 5881 | 35592 | 90.0% | 10.0% |
| Reddit | 67026 | 333866 | 94.2% | 5.8% |

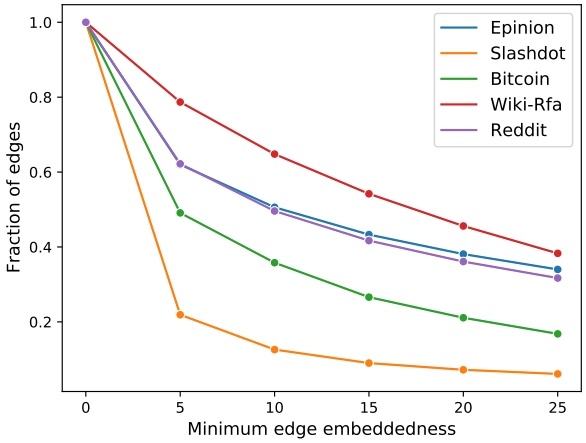



Figure 3: Fraction of edges under different embeddedness levels

## 5.2 Metrics

The following metrics are used to evaluate the model performance.

- Accuracy: the percentage of correct predictions. A good model should have high accuracy.
- False positive rate: the ratio between the number of negative edges wrongly categorized positive and the total number of negative edges. All the five networks have a extremely large proportion of positive edges, so the prediction accuracy can still reach very high at the cost of high false positive rate. Thus as suggested in [1], we also use the false positive rate as an indicator, which reflects the model's ability to reject the incorrect classification. Therefore we expect the model to have a low false positive rate while achieving high accuracy.

# 6. RESULTS

From table 1, we could see that the positive edges in the five datasets account for a overwhelming large of proportion, which means the random guessing could yield a high prediction accuracy. We will use this random baseline as comparison to see if the model could improve the prediction accuracy.

Table 2 illustrates the accuracy and false positive rate of the five datasets with zero-embeddedness. Overall, the model with triad-based features achieves better performance than the random baseline on the Epinion, Slashdot, Wiki-Rfa and Bitcoin datasets, while it performs worse on the Reddit dataset. However, the false positive rates on the five datasets are high especially for the Reddit data. By contrast, the model with 4-cycle obtains achieves very high accuracy and low false positive rate and outperforms that with triad-based features.

Table 2: Accuracy and false positive rate of the models

| | Accuracy | | False positive rate | |
|---|---|---|---|---|
| | 3-cycle | 4-cycle | 3-cycle | 4-cycle |
| Epinions | 0.924 | 0.993 | 0.460 | 0.039 |
| Slashdot | 0.851 | 1.000 | 0.530 | 0.002 |
| Wiki-Rfa | 0.848 | 0.998 | 0.498 | 0.005 |
| Bitcoin | 0.937 | 0.999 | 0.501 | 0.004 |
| Reddit | 0.940 | 1.000 | 0.987 | 0.005 |

Figure 4 and figure 5 correspond to the classification accuracy and false positive rate of the five datasets for the two types of k-cycle features, and for different minimum edge embeddedness levels. We will combine them to explain the results on the five datasets.

## 6.1 Degree and triad based features

Let us first focus on the performance of the model with the first feature combination: degree and triad based features.

- For the Epinion and Slashdot datasets, the model firstly performs better and then tend to be flat with an increasing of embeddedness levels, as expected. Greater embeddedness could provide a higher proportion of triangle information, thus the triad-features will work more effectively in such case.
- As the embeddedness level increases, the prediction accuracy on the wiki-Rfa dataset grows slowly, and the false positive rate even slightly increases. As explained in the paper [5], the election of Wikipedia administrator are affected by many factors, like the historical activities of candidates. Therefore it is insufficient to infer the preference of voters to the candidates only from the simple network structures. Conversely, a larger embeddedness level means the number of samples for training will become less, which might have a negative effect on the performance.
- It is noteworthy that the tendency and specific metric values of the above three datasets are close to the results obtained in [1]. The minor difference might be the result of the data evolution and different parameter selections.
- The Bitcoin dataset is a new dataset that is not experimented in the original paper. As the embeddedness levels become larger, the model has similar tendency with Epinion and Slashdot in terms of accuracy, but the false positive rate firstly decreases and

then has a slight increase. First it indicates that the triad-based features are very helpful to understand the trust relationships between the bitcoin users. On the other hand, It could not be ignored that the bitcoin dataset has the smallest number of edges among the five datasets and the training samples will become less with a larger embeddedness level, which might deteriorate the model performance.

- The Reddit dataset is another new dataset. The model performs worse than the random baseline, and the increase of embeddedness levels cannot improve the model and even worsen the results. We guess there are two mains reasons. First, the Reddit dataset has the largest proportion of positive edges among the five datasets. Such serious imbalance in the edge sign would bias the classifier to predict positive edges. This is also reflected in the high false positive rate. Another reason might be that the Reddit dataset is more complicated than the others and the sentiment between the subreddits is closely related to the post contents. The triangle structures are not capable of representing the relationships between the subreddits.

## 6.2 Incorporating quadrangle features

After adding the quadrangle based features, we found the model achieves nearly perfect classification on all five datasets. The results are surprisingly well and also better than the results presented in [1]. One possible reason is that the original paper reduces the number of k-cycle features by ignoring the directions for simplifying the computation complexity. That means the number of 4-cycle features will reduce from 64 to 8 if we treat it as an undirected network. In our work we still extract all cycle types from the directed network. The difference in the results also indicates that the reduction to the undirected network would lose much useful information related with the sign relationships between nodes.

# 7. CONCLUSION

In this project, we study the edge sign prediction problem and reproduce the methods from Leskovec et al.[5] and Chiang et al. [1] to tackle the problem. We evaluate the models on five different datasets.

From the results we can see that the quad type features can better explain the features of data than the triad types features. The results for quad type features reach nearly 100% accuracy while the false positive rate are near 0 which indicate the almost perfect performance through extracting quad type features. With respect to the performance of triad type features, except for Reddit dataset, the accuracy for other four datasets experience different degrees of enhancement with the higher embeddedness levels. In conclusion, training the logistic regression model by applying quad type features will obtain better results than applying triad type features.

The future improvement for this project can be in two directions. Firstly, it might be a good direction to apply the deep learning neural networks for link sign problem. Secondly, as we discussed before, the link sign in the Reddit data depend on many factors. We could combine the signed network structure with the post content as features into the model and see if they will enhance the performance of the model. In addition, we found there are many replicated edges in the Reddit data, so it might be necessary to weight the edges of signed networks.

# 8. REFERENCES


[1] K.-Y. Chiang, N. Natarajan, A. Tewari, and I. S. Dhillon. Exploiting longer cycles for link prediction in signed networks. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1157–1162, 2011.

[2] W. Goddard. Graphs and matrices, 2016.

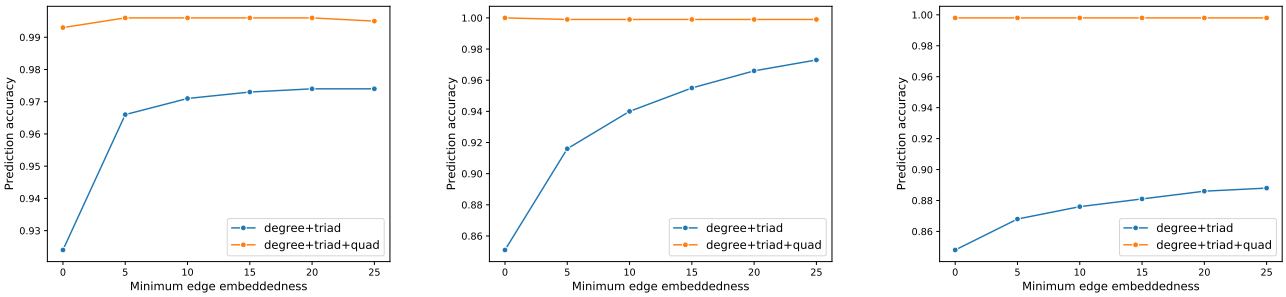
[3] C.-J. Hsieh, K.-Y. Chiang, and I. S. Dhillon. Low rank modeling of signed networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 507–515, 2012.

[4] A. Javari and M. Jalili. Cluster-based collaborative filtering for sign prediction in social networks with positive and negative links. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(2):1–19, 2014.

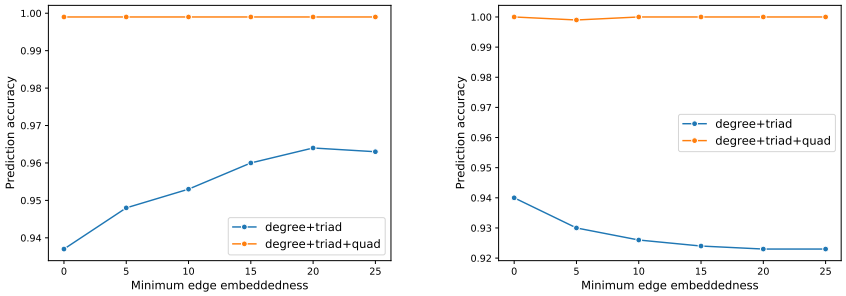
[5] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650, 2010.

[6] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[7] T. Zhang, H. Jiang, Z. Bao, and Y. Zhang. Characterization and edge sign prediction in signed networks. *Journal of Industrial and Intelligent Information Vol*, 1(1), 2013.

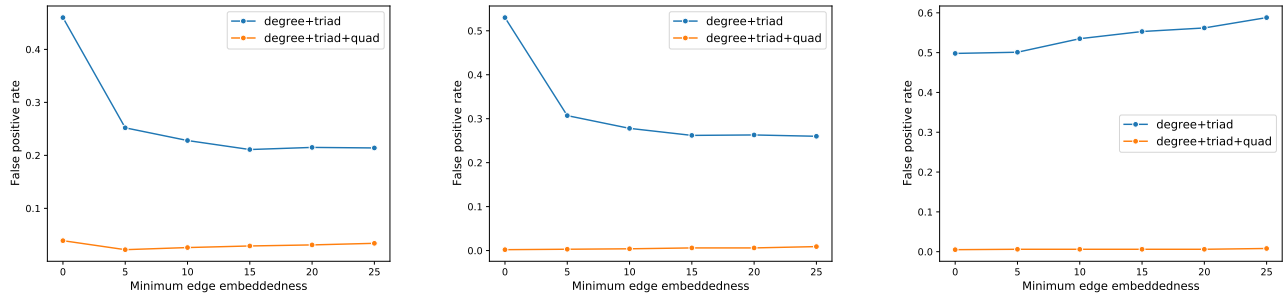



(a) Epinion (b) Slashdot (c) Wiki-Rfa

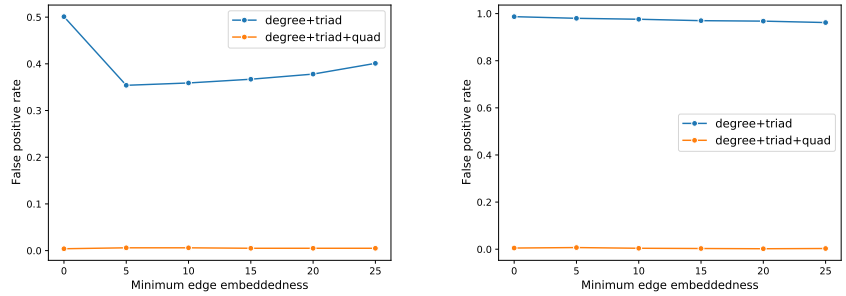



(d) Bitcoin (e) Reddit

Figure 4: Accuracy measures of the models with two feature combinations respectively: (1)degree+triad; (2) degree+triad+quad. The plots show the accuracy of the two models for different minimum edge embeddedness on the five datasets.




(a) Epinion (b) Slashdot (c) Wiki-Rfa




(d) Bitcoin (e) Reddit

Figure 5: False positive rate of the models with two feature combinations respectively: (1)degree+triad; (2) degree+triad+quad. The plots show the false positive rate of the two models for different minimum edge embeddedness on the five datasets.