Article

Spam Classification Based on Signed Network Analysis

Sihyun Jeong ¹ and Kyu-haeng Lee ^{2,*}

¹ Department of Computer Science and Engineering, Seoul National University, Seoul 08826, Korea; sihyunj@snu.ac.kr

² Department of Mobile Systems Engineering, Dankook University, Gyeonggi 16890, Korea

* Correspondence: kyuhaeng.lee@dankook.ac.kr

Received: 7 October 2020; Accepted: 11 December 2020; Published: 15 December 2020

Abstract: Online social networking services have become the most important information-sharing medium of modern society due to several merits, such as creating opportunities to broaden social relations, easy and instant communication, and fast data propagation. These advantages, however, are being abused by malicious users to disseminate unsolicited spam messages, causing great harm to both users and service providers. To address this problem, numerous spam detection methods utilizing various spam characteristics have been proposed, but most of them suffer from several limitations. Using individual behaviors and the content of messages for spam classification has been revealed to have bounded performance, since attackers can easily fake them. Instead, exploitation of social-network-related features has been highlighted as an alternative solution, but recent spam attacks can adroitly avoid these methods by controlling their ranking through various forms of attack. In this paper, we delineate a signed-network-analysis-based spam classification method. Our key hypothesis is that the edge signs are highly likely to be determined by considering users' social relationships, so there will be a substantial difference between the edge sign patterns of spammers and that of non-spammers. To identify our hypothesis, we employ two social psychological theories for signed networks—structural balance theory and social status theory—and the concept of surprise is adopted to quantitatively analyze the given network according to these theories. These surprise measurements are then used as the main features for spam classification. In addition, we develop a graph-converting method for applying our scheme to unsigned networks. Extensive experimental results with Twitter and Epinions datasets show that the proposed scheme obtains significant classification performance improvement compared to conventional schemes.

Keywords: online social network; security; spam; signed network; status theory; balance theory

1. Introduction

Recently, online social networking services, such as Twitter, Facebook, and Weibo, have become the most important means of information sharing in modern society. A great deal of news is quickly spread to unspecified people through the services, and they can easily chat or share their daily lives with their friends, colleagues, and family anywhere and at any time. Social media platforms are being used for political processes and activities, and online social media advertising is now considered to be one of the most effective marketing strategies. Furthermore, additional offered functionalities, such as entertainment, shopping, banking, and games, accelerate the growth of online social networking services. According to research, Facebook has 2.60 billion monthly active users and also has 1.73 billion users that are visiting the social networking site on a daily basis [1].

As the use of online social networking services increases and their influence becomes more significant, a number of misuse and abuse cases have been continuously reported. A typical example

is spam. For various purposes, malicious attackers try to disseminate unsolicited messages, such as indiscriminate advertisements, phishing links, and fake news, by taking advantage of online social networking services. Spam attacks hinder the development of better social networking services and degrade their reliability and credibility, causing great damage to both users and service providers: Users annoyed with spam may leave the platform, and the tremendous cost of preventing spam attacks puts enormous pressure on service providers.

Numerous schemes to tackle spam attacks have been proposed so far. Initial attempts rely on content-based detection methods. They capture vocabulary or phrases appearing frequently in spam messages [2–4], and they have helped detect a significant amount of spam to this day. Detection performance can be further improved by utilizing more features, such as language, usage time, location, and even friend selection criteria [2–5]. However, these methods, which mainly use the content or behavior of the message, quickly become less effective as the attacks evolve to be more complex and multifaceted. Instead, exploiting social-network-related features has recently been highlighted, since attackers cannot easily imitate social relationships. Many experimental results reveal that the network features obtained through social network analysis, such as node degree, clustering coefficient, and rank, can be used effectively as a basis for spam classification [6–11]. In particular, PageRank [12–14] and Hyperlink-Induced Topic Search (HITS) [15] are extensively exploited here, since they are designed to reflect these social network properties well. In these methods, nodes with low PageRank or HITS scores have a high probability of being classified as spammers.

Unfortunately, recent spam attacks can adroitly avoid those spam detection approaches by controlling their ranking through link farming and Spamdexing [8,10,16,17]. For handling these attacks, the signs of edges could be significantly helpful, since they help us understand not only the relationship between two connected nodes, but also the reasoning of the link-forming process. Nevertheless, they are still being treated as unimportant and are not fully exploited for spam classification studies. In addition, since most current social networking services allow users to express their emotions and opinions in various ways, such as with likes/dislikes, comments, or ratings, we can obtain the sign attributes from the data without difficulty.

In this paper, we present a signed-network-analysis-based spam classification method. Our key hypothesis is that the edge signs are highly likely to be determined by considering users' social relationships, so there will be a substantial difference between the edge sign patterns of spammers and those of non-spammers (i.e., normal users). To test our hypothesis, we employ two social psychological theories for signed networks—structural balance theory and social status theory—and the concept of surprise is adopted to quantitatively analyze the given network according to these theories. These surprise measurements are then used as the main features for spam classification. Through extensive experiments with a Twitter dataset, we observe that the proposed spam classification scheme achieves an accuracy of 93%, a precision of 95%, and a recall of 91%. These metrics are higher than those of conventional spam detection schemes. Since the proposed scheme basically utilizes edge sign information, we cannot directly apply it to unsigned networks. To handle this issue, we additionally develop a graph-converting method. Through the analysis of the Epinions dataset, we discover the fact that the edge signs have strong correlations with PageRank and Conductance scores, and based on that, we propose the assignment of positive and negative signs to the edges according to those scores. Extensive real-world dataset experiments verify that the proposed graph-converting method works properly for unsigned networks.

The main contributions of this paper are as follows:

- (1) To the best of our knowledge, the proposed scheme is the first spam classification method that utilizes the structural balance theory and social status theory. We verify the feasibility of the proposed scheme through extensive experiments with Twitter and Epinions datasets and a comparison with five social-network-analysis-based spam detection approaches.

- (2) For applying our scheme to unsigned networks, we develop a graph-converting method. To do this, we utilize the fact that the edge signs have strong correlations with PageRank and Conductance scores from the analysis of the Epinions dataset.

The remainder of this paper is organized as follows. In Section 2, we summarize several spam classification schemes related to this paper. Section 3 provides background information on the structural balance theory and social status theory. The feature validation is described in Section 4, and Section 5 presents the results of a performance evaluation. Finally, we conclude this paper in Section 6.

2. Related Work

Most spam detection studies consider spam detection as a classification problem. They focus on identifying the appropriate data features or best combinations of them that can distinguish spam/non-spam or a spammer/non-spammer. One simple way to detect spam is to examine the text of the message: a spam message is likely to be in a format that may feel awkward to people, or its contents may be meaningless or even non-relevant to the recipients [2,3]. A template-matching scheme is proposed to detect spam messages by comparing the sentence structure of a given message with that of ground-truth, non-spam messages [18]. Since spam messages may contain a hashtag or URL for quick spreading, checking a fraction of them has been widely used for several proposals [2,4]. Juan et al. examine the use of language in the topics and pages linked from a tweet [3], and Yardi analyzes trending topics on Twitter to understand spammers' strategic behavioral patterns [5]. In addition to the text itself, meta information, such as the user profile, activity logs, and timestamps, has been also used for spam filtering [2–5]. SynchroTrap uses time-stamped user actions to detect malicious account groups that generate “like spam” on Facebook or “follow spam” on Instagram [19]. VOLTINE [20] utilizes an inter-arrival time distribution of patterns of how normal users write reviews. The patterns of webpage click sessions, such as traffic moderateness, target synchronicity, and temporal synchronicity, are used to detect click spam attacks [21,22]. Most of these approaches are intuitive and simple to implement; however, they become rapidly ineffective, as attacks are evolving to be cleverer and more multifaceted. To capture sophisticated online attacks, many representation-learning-based spam detection approaches use both relational and semantic features [23–25].

Several network properties accumulated while using online social networking services have been highlighted as alternative features for spam classification. Due to the fact that attackers cannot easily imitate social relationships, spam detection approaches based on social network analysis are considered to be more robust than the previously described schemes. One simple yet effective realization method is to use node ranking: The rankings of nodes are determined based on combinations of several network properties, and spammers and non-spammers are classified according to these rankings. PageRank [12–14] and HITS [15] are extensively exploited here, since they are designed to reflect these social network properties well. Unfortunately, spammers can adroitly avoid such spam detection approaches through various forms of attack, such as link farming and Spamdexing [8,10,16,17]. To address this problem, CatchSync [16] uses the observation that anomalies in online social networks tend to follow (or make links to) users whose node degrees and HITS values are similar to each other. Similarly, recent research [26,27] utilized spammers' synchronicity by analyzing activity burstness. SybilRank [10] utilizes the fact that Sybils (i.e., spammers) have a disproportionately small number of connections to non-Sybil users. It provides a ranking algorithm that penalizes Sybils and link-farming activities. Network Footprint Score (NFS) [17] captures social campaigners by quantifying the likelihood of spam campaign targets based on their PageRank scores. Li et al. use a graph propagation algorithm based on the answerer-channel bipartite graph to detect spam answers in Community Question Answering sites [11]. Integro [9] is an optimized random walk-based ranking algorithm for fake account detection. SybilEdge [28] uses friend request acceptance/rejection activities for the detection of new fake accounts on social networks. Triad Significance Profile (TSP)-Filtering [29] exploits the structural differences between the frequencies of certain triad motifs for spam filtering. These approaches resemble

our approach in that social relationships are taken into account for spam classification, but they have the limitation that their applications are restricted to only undirected or unsigned networks.

3. A Primer on Social Psychological Theories for Signed Networks

In this section, we introduce two theories of social psychology—structural balance theory and social status theory—as background information. These theories have been advanced for exploring different aspects of signed social networks, where each edge is encoded with a certain sign value: positive or negative. Edge signs can provide more detailed information about the relationship between two connected nodes and the reasoning behind the link-forming process, and are therefore very meaningful. To analyze how well a given network can be explained by each theory, we employ a metric called surprise, which was proposed by Leskovec et al. [30]. Basically, a surprise value is computed for a certain type of motif, as shown in Figures 1 and 2, and shows how much the target motif or a certain link of the motif appears in the actual network compared to our expectation. Simply speaking, a positive surprise value means that the corresponding motif or a certain link of that motif appears more in the actual network than expected, and vice versa for negative surprise. Here, the expectation is represented differently in the two theories: For the structural balance theory, the randomly shuffled network is used as a baseline, while for the social status theory, the overall fraction of positive edges that a node expects to create or receive in the target motif is used as a baseline. In this paper, we use these surprise values as key network features for classifying spammers and non-spammers. In the following, we review the two theories and the concept of surprise in more detail.

Figure 1. Four triad types for unsigned networks. The index of each triad is the number of positive signs of that triad.

Figure 2. Sixteen triad types for signed directed networks. We are interested in predicting the sign of the red link (i.e., A–B link) considering the signs of the A–X link and B–X link.

3.1. Structural Balance Theory

Structural balance theory, first formulated by Heider [31], can be simply explained with the following two claims: (1) “The friend of my friend is my friend” and (2) “the enemy of my friend is my enemy” (“the enemy of my enemy is my friend” and “the friend of my enemy is my enemy” can be regarded as the same notion). In Figure 1, triad types T_3 and T_1 correspond to the first and second claim, respectively. According to this theory, types T_1 and T_3 are considered to be balanced, while types T_0 and T_2 are considered to be unbalanced, which implies that conceptually, types T_1 and T_3 are more plausible in the real world than the other types.

The surprise value for a certain motif T_i is defined with the following formula [30]:

$$s_b(T_i) = \frac{T_i - E[T_i]}{\sqrt{\Delta p_0(T_i)(1 - p_0(T_i))}}, \quad (1)$$

where $E[T_i]$ is the expected number of type T_i , and Δ is the total number of triads in the network. $\Delta p_0(T_i)$ is the expected fraction of triads that are of type T_i after shuffling the signs of all edges in the graph.

As explained earlier, $s_b(T_i)$ indicates how much the frequency of appearance of the triad type T_i differs compared to the randomly shuffled network. Note that the fraction of positive edges is kept during the shuffling. According to research, many existing online social network services have high positive surprise values for type T_3 and low surprise values for type T_2 [30]. One thing to note is that the structural balance theory is intended for undirected graphs, as shown in Figure 1. Of course, for directed graphs, we can apply the structural balance theory by simply ignoring the direction of the edge, which may cause a decrease in classification performance due to the information loss for edge directionality.

3.2. Social Status Theory

Social status theory explains that the sign of an edge is determined by the status difference between nodes. For example, the link created from node x to node y would be positive if node x considers node y to have a higher status than itself. Consider the edge sign prediction problem illustrated in Figure 2. Here, our goal is to predict the unknown sign of the edge from node A to node B (simply denoted as the A – B edge), i.e., the red dotted link, using the A – X edge and the B – X edge. Unlike the surprise measurement of structural balance theory, the surprise measurement used with social status theory is conducted separately for the edge initiator and the edge destination, since it considers directed networks, and these measurements are called the generative surprise and the receptive surprise, respectively. According to a study by Leskovec [30], the generative surprise (denoted as $s_g(T_i)$) is computed by the number of standard deviations by which the actual number of positive A – B edges in the data differs from the expected positive numbers created by the baseline. Here, the baseline is defined as the overall fraction of positive edges that a user expects to create. Consider type T_{12} in Figure 2. For node A , since it receives a positive sign from node X , it might think that it has a higher status than others. In this case, according to social status theory, it will be likely to give negative signs to other nodes (e.g., node B), but if the actual sign of the A – B edge turns out to be positive, then the generative surprise value will be high. Receptive surprise is a concept similar to generative surprise, except that it is computed from the viewpoint of the edge destination. From the perspective of node B , since it gives a negative sign to node X , it might expect that it will receive positive signs from others (e.g., node A). In this case, if the actual sign of the A – B edge is positive, then the receptive surprise becomes low. From this example, we can see that for the same motif, there could be different interpretations depending on the viewpoint. For another example, in the case of type T_{11} , the theory considers the status of node X to be higher than that of node B , but lower than that of node A , and as a result, it predicts that the red arrow is negative. Interestingly, this result is completely opposite to the result of the structural balance theory. Recall that in the balance theory, this edge should be positive to be balanced (T_3 in Figure 1). As can be

seen, both theories could result in different views of the same network, but in this paper, we do not discuss the subtle, differing interpretations that these two theories can make about the same network, since this is beyond the scope of this paper. We do, however, note that they are both still effective as key factors for discrimination between spammers and non-spammers, which will be shown later.

4. Feature Extraction and Validation

To investigate how well our hypothesis fits to actual data, we measure the aforementioned surprise values for spammers and non-spammers in a Twitter dataset [8,32]. The basic statistics of the dataset are given in Table 1. Note that this dataset includes the information about labeled spammers, which we can use as ground-truth data. Since the two theories that we adopt in this paper are intended for signed networks, if the given network does not have sign information, then we cannot directly apply our scheme. Note that the Twitter dataset has no sign information. To handle this issue, we develop a graph-converting method, which will be explained in the following subsection.

Table 1. Statistics of the Twitter dataset.

	Twitter
The number of users	54,981,152
The number of spammers	41,352
The number of following links	1,963,263,821

4.1. Coping with Unsigned Networks

In order to convert the given unsigned network into a signed network, we have to artificially assign signs to the edges of the network. Our idea is simple: Among all E edges, we select $\alpha \cdot E$ edges and assign them negative signs (and positive signs to the others). α is the fraction of negative edges, and we discuss how to select α in a later subsection. One challenging issue here is how to appropriately select those $\alpha \cdot E$ edges. The underlying motivation behind transforming into signed networks is that there exist correlations between the edge signs and social status of nodes. To investigate this relationship, we analyze the properties of the actual signed network using the Epinions dataset [30]. We choose two well-known metrics, PageRank and Conductance, since they are suitable for understanding the influence of a node or a cluster [13,14,33]. The basic statistics of the dataset are given in Table 2.

Table 2. Statistics of the Epinions dataset.

	Epinions
The number of nodes	131,828
The number of edges	841,372
The ratio of positive edges	85%
The ratio of negative edges	15%

4.1.1. PageRank Analysis

PageRank, which is a way to measure the importance of a node, has been considered one of the most widely used node influence measurement indices [13,34,35]. In this analysis, we find out whether there exists a correlation between the PageRank score of a node and its edge types through both node-level analysis and edge-level analysis. In the node-level analysis, for all nodes, we measure the ratio of edge types connected to each node. For example, in the example of Figure 3, Node A is connected to three positive edges, two of which are incoming edges and the other is an outgoing edge. Therefore, the ratio of positive incoming edges of Node A is $2/3$. Similarly, for the negative edges, the ratio of negative incoming edges of Node A is $1/2$. Note that the ratio values for positive edges and negative edges are calculated separately.

Figure 3. An example of signed networks. A node may be connected to several nodes through different types of edges.

Figure 4 shows the measurement results. From Figure 4a, we can see that high ratio values of positive incoming edges are observed more in nodes with high PageRank scores than in those with low PageRank scores. In the case of negative edges (Figure 4b), though the result is less pronounced than in the case of positive edges (Figure 4a), we can see that nodes with high PageRank scores have more outgoing edges. A similar result is observed in average PageRank measurements in Figure 4c; for positive edges, the PageRank score of the source node is lower than of that of the destination node, and vice versa for negative edges.

Figure 4. PageRank measurements at the node level. Nodes with higher PageRank scores have more positive incoming edges and negative outgoing edges than those with lower scores.

In the edge-level analysis, we focus on the difference of PageRank scores between the edge initiator and the edge destination. Table 3 shows a breakdown of the number of edges according to the difference of PageRank scores. For each edge, the difference, d , is computed by subtracting the PageRank score of the destination node from that of the source node. As can be seen from the table, the higher the destination PageRank score is than the source PageRank score (i.e., $d < 0$),

the more positive edges are observed (64.67%). In the opposite case (i.e., $d > 0$), more negative edges are observed (64.99%). The above observations are consistent with the social status theory, where a lower-status node (i.e., lower PageRank score) is more likely to give a positive sign to a higher-status node (i.e., higher PageRank score).

Table 3. The number of edges according to PageRank differences.

	$d < 0$	$d > 0$	$d = 0$
Positive Edge	464,119 (64.67%)	252,135 (35.13%)	1412 (0.20%)
Negative Edge	43,261 (34.97%)	80,396 (64.99%)	49 (0.04%)

4.1.2. Conductance Analysis

We can see a similar tendency from the clustering analysis in Figure 5. In this experiment, using the Louvain clustering method [36], we first generate clusters and then measure the Conductance for each cluster that an edge belongs to. The Conductance gives a score for a cluster by considering its internal and external connectivity [37]. As the name suggests, higher Conductance means higher separability of the cluster.

Figure 5. Average Conductance measurements of the edges in the Epinions dataset. A similar result to that of Figure 4 is observed in the Conductance measurement. For positive edges, the Conductance score of the source node is higher than of that of the destination node, and vice versa for negative edges.

The Conductance of cluster S is computed as the following:

$$f(S) = \frac{c_s}{2m_s + c_s}, \quad (2)$$

where c_s is the number of edges on the boundary of S , and m_s is the number of edges in S .

From the analysis result in Figure 5, we can see that for positive edges, the Conductance score of the source node is higher than of that of the destination node, and vice versa for negative edges. When comparing this result with the result of PageRank in Figure 4c, we can see that both have similar views on forming edge signs: the higher the internal connectivity and the lower the external separability a node has, the higher its PageRank score.

4.1.3. Graph-Converting Method

Based on the above observations, we take both PageRank and Conductance into account to select $\alpha \cdot E$ edges. For a randomly selected edge, we assign a negative sign to it if the following two conditions are satisfied: (1) The PageRank of the edge initiator is higher than that of the edge destination, and (2) the Conductance of the edge initiator is lower than that of the edge destination. This process continues until we have $\alpha \cdot E$ negative edges. For setting the α value, we refer to the statistics of real-world datasets.

According to research, many online social network services, such as Epinions, Wikipedia, and Slashdot, have a negative sign over about 20% of all links [30], and based on this result, we set α to 0.2.

4.2. Validation

As mentioned before, to validate our hypothesis, we measure balance surprise and status surprise for spammers and non-spammers in a Twitter dataset. To do this, we randomly select 1000 spammers and 1000 non-spammers (i.e., 2000 nodes in total), and then construct an ego-network for each. After that, we measure the average balance and status surprise value for each network.

Figure 6 shows the comparison of the structural balance surprise values of spammers' networks and non-spammers' networks. We can see that the characteristics of spammers and non-spammers are markedly different in terms of surprise. In particular, in the case of normal users, type T_3 is overrepresented in the network, while type T_2 is underrepresented. Recall that this result is consistent with Heider's claim about the structural balance theory [31]. According to the theory, type T_1 and type T_3 are more plausible than other types in real-world social networks, which is clearly observed in the case of normal users. On the other hand, this pattern is not clearly shown for spammers. In particular, the balance surprise gap between types T_2 and T_3 is quite large in the case of normal users, but not in the case of spammers. Rather, we find a case inconsistent with the structural balance theory in the case of spammers; the balance surprise of type T_1 is negative, which cannot be easily explained by balance theory.

Figure 7a,b shows the generative surprise and receptive surprise for spammers and non-spammers, respectively. First, for the generative surprise result, we can see that both normal users and spammers have results consistent with the social status theory, i.e., the same surprise signs for all of the triads, but with different degrees of consistency (in particular, types T_2 , T_8 , and T_{10}). This difference appears more clearly in the receptive surprise result in Figure 7b. For most types of triads, we can see a large surprise gap between normal users and spammers. From these observations, we verify that spammers and non-spammers have different apparent network properties in terms of balance and status surprise, which shows the feasibility of using them as a basis for spam classification. All of these surprise metrics show the influence of social relationships on the edge signs. Therefore, the distinctive surprise patterns that appear in normal user cases imply that social relationships greatly affect the formation of the edge signs, as stated in our hypothesis.

Although these results indicate that the proposed graph-converting method is reasonable, for better verification, an in-depth investigation into the performance (e.g., accuracy) is required. We leave this as our future work.

Figure 6. Balance surprise. The variance of the balance surprise values of non-spammers' ego-networks is bigger than that of spammers' ego-networks.

(a) Generative surprise

(b) Receptive surprise

Figure 7. Generative surprise and receptive surprise of spammers' ego-networks and non-spammers' ego-networks. Spammers and non-spammers show different patterns of status surprise values. Similarly to the results of balance surprise, the variance of the status surprise values of non-spammers is bigger than that of spammers. Such a difference appears more clearly in the receptive surprise result.

5. Performance Evaluation

5.1. Settings

In this section, we investigate the performance of the proposed scheme by using the same Twitter dataset described in Table 1. We measure accuracy, precision, recall, false positive rate, and F1 score as performance metrics. For comparison, we implement the following four spam detection methods: SybleRank [10], NFS [17], CatchSync [16], and TSP-Filtering [29]. We use C4.5 [38] in Weka [39] as the classifier (note that it is known as J48 in Weka). The batch size is set as 100, and 10-fold cross-validation is used. Surprise measurements for triad types in Figures 1 and 2 are used as the main features for the classifier:

- Balance: using only the balance feature;
- Status: using only the status feature;
- Balance and Status: using both the balance and status features.

Table 4 shows the information gain values in the decision tree when both the balance and status features are used. In the table, BT and ST refer to the triad types in Figures 1 and 2, respectively. These results are sorted by gain value.

Table 4. Information gain in the decision tree.

Rank	Information Gain	Feature	Rank	Information Gain	Feature
1	0.639	ST9	11	0.522	BT2
2	0.613	ST1	12	0.507	ST8
3	0.609	BT3	13	0.436	ST14
4	0.605	T11	14	0.411	ST2
5	0.603	ST3	15	0.363	BT0
6	0.584	BT1	16	0.348	ST10
7	0.573	ST5	17	0.339	ST13
8	0.560	ST12	18	0.311	ST15
9	0.552	ST7	19	0.251	ST6
10	0.551	ST4	20	0.211	ST16

5.2. Results

Table 5 shows the comparison result. The node-ranking-based detection approaches, SybilRank and NFS, show poor spam detection performance for all indicators; they achieve detection accuracies of 28% and 57%, respectively, and F1 scores of less than 51%, while the other approaches achieve accuracies of more than 90%. The proposed schemes outperform both CatchSync and TSP in terms of accuracy and precision. They achieve an accuracy of more than 92% and a precision of almost 95%. For recall, however, the proposed schemes have lower scores than CatchSync and TSP. Although they show the highest recall performance, their false positive rates are close to 10%, implying that they aggressively classify spammers. On the other hand, the proposed schemes show more stable detection performance; they can achieve almost the same level of detection performance while maintaining a low false positive rate. TSP resembles our approach in that the triad patterns are utilized, but it considers only undirected networks, which leads to performance degradation.

Table 5. Performance comparison result.

	Accuracy	Precision	Recall	FPR	F1-Score
SybilRank	0.283	0.303	0.335	0.769	0.318
NFS	0.572	0.597	0.443	0.299	0.509
CatchSync	0.903	0.894	0.915	0.109	0.904
TSP	0.908	0.906	0.911	0.095	0.908
Balance	0.925	0.943	0.904	0.055	0.923
Status	0.929	0.949	0.907	0.049	0.927
Balance + Status	0.928	0.948	0.906	0.050	0.926

For the proposed method, we obtain the best performance result when we use only the social status features; when the balance feature is used, the classification performance is slightly degraded. This is because the balance feature ignores the information about the edge direction of the Twitter dataset, which results in a weak fit to the data. Nevertheless, we can see that it is still more effective for spam classification than conventional node-ranking-related features.

6. Conclusions and Future Work

In this paper, we present a spam classification method based on the structural balance theory and social status theory. Our key hypothesis is that the edge signs are highly likely to be determined by considering users' social relationships, so there will be a substantial difference between the edge sign patterns of spammers and those of non-spammers. Through Twitter data analysis, we can observe the apparent difference for spammers and non-spammers in terms of surprise, and we show that the proposed scheme obtains significant classification performance improvement compared to conventional schemes. It achieves an accuracy of 93%, a precision of 95%, and a recall of 91%. In addition, to apply our scheme to unsigned networks, we develop a graph-converting method.

From the analysis of the Epinions dataset, we discover the fact that the signs of edges have strong correlations with PageRank and Conductance scores, and based on that, we assign positive and negative signs to the edges. Extensive real-world dataset experiments verify that the proposed converting method works well for unsigned networks.

We believe that we can not only further improve the performance of the proposed scheme, but also broaden the scope of it to a variety of applications that can be modeled as signed networks. For example, in recommendation systems, the sign of an edge might be determined based on ratings. In social blogging or community sites, sentiment analysis of users' reviews and comments might be conducted for identifying attributes of nodes and links. The proposed scheme is basically orthogonal to conventional behavior-based or content-based spam detection technologies, and thus, they both complement each other. In particular, advanced classifiers with graph-embedding methods could be employed for achieving better classification performance. In the case that the given network data lack information about node and edge attributes, node and edge prediction methods through semi-supervised learning could be applied. We leave the further study of these issues for future work.

Author Contributions: Conceptualization, S.J.; methodology, S.J.; writing—review and editing, S.J. and K.-h.L.; funding acquisition, K.-h.L. All authors have read and agreed to the published version of the manuscript.

Funding: The present research was supported by the research fund of Dankook University in 2020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aboulhosn, S. Facebook Statistics. 2020. Available online: <https://sproutsocial.com/insights/facebook-stats-for-marketers> (accessed on 3 August 2020).
2. Benevenuto, F.; Magno, G.; Rodrigues, T.; Almeida, V. Detecting spammers on twitter. In Proceedings of the Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), Redmond, WA, USA, 13–14 July 2010.
3. Martinez-Romo, J.; Araujo, L. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Syst. Appl.* **2013**, *40*, 2992–3000. [CrossRef]
4. Egele, M.; Stringhini, G.; Kruegel, C.; Vigna, G. Towards Detecting Compromised Accounts on Social Networks. *IEEE Trans. Dependable Secur. Comput.* **2017**, *14*, 447–460. [CrossRef]
5. Yardi, S.; Romero, D.; Schoenebeck, G.; Boyd, D. Detecting Spam in a Twitter Network. *First Monday* **2010**, *15*. [CrossRef]
6. Viswanath, B.; Bashir, M.A.; Crovella, M.; Guha, S.; Gummadi, K.P.; Krishnamurthy, B.; Mislove, A. Towards detecting anomalous user behavior in online social networks. In Proceedings of the 23rd USENIX Security Symposium (USENIX Security 14), San Diego, CA, USA, 20–22 August 2014; pp. 223–238.
7. Stringhini, G.; Wang, G.; Egele, M.; Kruegel, C.; Vigna, G.; Zheng, H.; Zhao, B.Y. Follow the green: Growth and dynamics in twitter follower markets. In Proceedings of the 2013 Conference on Internet Measurement Conference, Barcelona, Spain, 23–25 October 2013; pp. 163–176.
8. Ghosh, S.; Viswanath, B.; Kooti, F.; Sharma, N.K.; Korlam, G.; Benevenuto, F.; Ganguly, N.; Gummadi, K.P. Understanding and combating link farming in the twitter social network. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; pp. 61–70.
9. Boshmaf, Y.; Logothetis, D.; Siganos, G.; Lería, J.; Lorenzo, J.; Ripeanu, M.; Beznosov, K. *Integro: Leveraging Victim Prediction for Robust Fake Account Detection in OSNs*; NDSS, San Diego, CA, USA, 2015; Volume 15, pp. 8–11.
10. Cao, Q.; Sirivianos, M.; Yang, X.; Pregueiro, T. Aiding the detection of fake accounts in large scale social online services. In Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12), San Jose, CA, USA, 25–27 April 2012; pp. 197–210.
11. Li, X.; Liu, Y.; Zhang, M.; Ma, S.; Zhu, X.; Sun, J. *Detecting Promotion Campaigns in Community Question Answering*; IJCAI: Buenos Aires, Argentina, 2015; Volume 15, pp. 2348–2354.
12. Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*; Technical Report; Stanford InfoLab: Stanford, CA, USA, 1999.

13. Chen, R.; Hua, Q.; Wang, B.; Zheng, M.; Guan, W.; Ji, X.; Gao, Q.; Kong, X. A novel social recommendation method fusing user's social status and homophily based on matrix factorization techniques. *IEEE Access* **2019**, *7*, 18783–18798. [\[CrossRef\]](#)
14. Yin, X.; Hu, X.; Chen, Y.; Yuan, X.; Li, B. Signed-PageRank: An Efficient Influence Maximization Framework for Signed Social Networks. *IEEE Trans. Knowl. Data Eng.* **2019**. [\[CrossRef\]](#)
15. Kleinberg, J.M. Authoritative sources in a hyperlinked environment. *J. ACM (JACM)* **1999**, *46*, 604–632. [\[CrossRef\]](#)
16. Jiang, M.; Cui, P.; Beutel, A.; Faloutsos, C.; Yang, S. Catchsync: Catching synchronized behavior in large directed graphs. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014; pp. 941–950.
17. Ye, J.; Akoglu, L. Discovering opinion spammer groups by network footprints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 267–282.
18. Gao, H.; Yang, Y.; Bu, K.; Chen, Y.; Downey, D.; Lee, K.; Choudhary, A. Spam ain't as diverse as it seems: Throttling OSN spam with templates underneath. In Proceedings of the 30th Annual Computer Security Applications Conference, New Orleans, LA, USA, 8–12 December 2014; pp. 76–85.
19. Cao, Q.; Yang, X.; Yu, J.; Palow, C. Uncovering large groups of active malicious accounts in online social networks. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, 3–7 November 2014; pp. 477–488.
20. Chino, D.Y.; Costa, A.F.; Traina, A.J.; Faloutsos, C. VolTime: Unsupervised Anomaly Detection on Users' Online Activity Volume. In Proceedings of the 2017 SIAM International Conference on Data Mining, SIAM, Houston, TX, USA, 27–29 April 2017; pp. 108–116.
21. Li, X.; Zhang, M.; Liu, Y.; Ma, S.; Jin, Y.; Ru, L. Search engine click spam detection based on bipartite graph propagation. In Proceedings of the 7th ACM International Conference on Web Search and Data Mining, New York, NY, USA, 24–28 February 2014; pp. 93–102.
22. Tian, T.; Zhu, J.; Xia, F.; Zhuang, X.; Zhang, T. Crowd fraud detection in internet advertising. In Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 1100–1110.
23. Maity, S.K.; KC, S.; Mukherjee, A. Spam2vec: Learning biased embeddings for spam detection in twitter. In Proceedings of the The Web Conference 2018, Lyon, France, 23–27 April 2018; pp. 63–64.
24. Yuan, C.; Zhou, W.; Ma, Q.; Lv, S.; Han, J.; Hu, S. Learning review representations from user and product level information for spam detection. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019; pp. 1444–1449.
25. Liu, Z.; Dou, Y.; Yu, P.S.; Deng, Y.; Peng, H. Alleviating the Inconsistency Problem of Applying Graph Neural Network to Fraud Detection. *arXiv* **2020**, arXiv:2005.00625.
26. Ji, S.j.; Zhang, Q.; Li, J.; Chiu, D.K.; Xu, S.; Yi, L.; Gong, M. A burst-based unsupervised method for detecting review spammer groups. *Inf. Sci.* **2020**, *536*, 454–469. [\[CrossRef\]](#)
27. Li, H.; Fei, G.; Wang, S.; Liu, B.; Shao, W.; Mukherjee, A.; Shao, J. Bimodal distribution and co-bursting in review spam detection. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 1063–1072.
28. Breuer, A.; Eilat, R.; Weinsberg, U. Friend or Faux: Graph-Based Early Detection of Fake Accounts on Social Networks. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 1287–1297.
29. Jeong, S.; Noh, G.; Oh, H.; Kim, C.k. Follow spam detection based on cascaded social information. *Inf. Sci.* **2016**, *369*, 481–499. [\[CrossRef\]](#)
30. Leskovec, J.; Huttenlocher, D.; Kleinberg, J. Signed networks in social media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, GA, USA, 10–15 April 2010; pp. 1361–1370.
31. Heider, F. Social perception and phenomenal causality. *Psychol. Rev.* **1944**, *51*, 358. [\[CrossRef\]](#)
32. Cha, M.; Haddadi, H.; Benevenuto, F.; Gummadi, P.K. Measuring user influence in twitter: The million follower fallacy. *ICWSM* **2010**, *10*, 30.
33. Zhou, D.; Zhang, S.; Yildirim, M.Y.; Alcorn, S.; Tong, H.; Davulcu, H.; He, J. A Local Algorithm for Structure-Preserving Graph Cut. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 13–17 August 2017; pp. 655–664.

34. Riquelme, F.; González-Cantergiani, P. Measuring user influence on Twitter: A survey. *Inf. Process. Manag.* **2016**, *52*, 949–975. [[CrossRef](#)]
35. Rosa, H.; Carvalho, J.P.; Astudillo, R.; Batista, F. Detecting user influence in twitter: Pagerank vs. katz, a case study. In Proceedings of the Seventh European Symposium on Computational Intelligence and Mathematics, Cádiz, Spain, 7–10 October 2015, pp. 7–10.
36. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [[CrossRef](#)]
37. Chung, F.R.; Graham, F.C. *Spectral Graph Theory*; American Mathematical Soc.: Providence, RI, USA 1997.
38. Quinlan, J. *C4. 5: Programs for Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2014.
39. Witten, I.H.; Frank, E.; Hall, M.A. Weka 3. 2011. Available online: <https://www.cs.waikato.ac.nz/ml/weka/> (accessed on 1 September 2020).

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).