공학박사학위논문

개인 사회망 네트워크 분석 기반 온라인 사회 공격자 탐지

Online Social Attacker Detection based on Ego-network Analysis

2020 년 2 월

서울대학교 대학원

컴퓨터공학부

정 시 현

개인 사회망 네트워크 분석 기반 온라인 사회 공격자 탐지

Online Social Attacker Detection based on Ego-network Analysis

지도교수 김 종 권

이 논문을 공학박사 학위논문으로 제출함

2020 년 2 월

서울대학교 대학원

컴퓨터공학부

정 시 현

정시현의 공학박사 학위논문을 인준함

2020 년 2 월

위 원 장 : 이 상 구 (인)

부위원장 : 김 종 권 (인)

위 원 : 권 태 경 (인)

위 원 : 강 유 (인)

위 원 : 최 재 혁 (인)

Abstract

Online Social Attacker Detections based on Ego-network Analysis

Sihyun Jeong
Department of Computer Science & Engineering
The Graduate School
Seoul National University

In the last decade we have witnessed the explosive growth of online social networking services (SNSs) such as Facebook, Twitter, Weibo and LinkedIn. While SNSs provide diverse benefits – for example, fostering inter-personal relationships, community formations and news propagation, they also attracted uninvited nuiance. Spammers abuse SNSs as vehicles to spread spams rapidly and widely. Spams, unsolicited or inappropriate messages, significantly impair the credibility and reliability of services. Therefore, detecting spammers has become an urgent and critical issue in SNSs. This paper deals with spamming in Twitter and Weibo. Instead of spreading annoying messages to the public, a spammer follows (subscribes to) normal users, and followed a normal user. Sometimes a spammer makes link farm to increase target account's explicit influence. Based on the assumption that the online relationships of spammers are different from those of normal users, I proposed classification schemes that detect online social attackers including spammers. I firstly focused on ego-network social relations and devised two features, structural features based on Triad Significance Profile (TSP) and relational semantic features based on hierarchical homophily in an ego-network. Experiments on real Twitter and

Weibo datasets demonstrated that the proposed approach is very practical. The proposed features are scalable because instead of analyzing the whole network, they inspect user-centered ego-networks. My performance study showed that proposed methods yield significantly better performance than prior scheme in terms of true positives and false positives.

Keywords: Online Social Network, Spam, Ego-network, Homophily, Status, Anomaly detection, User feature analysis

Student number : 2014-21789

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The use of social networking services (SNSs) continues to grow exponentially with the widespread adoption of smart devices such as smart phones, smart pads, smart watches, and so on. SNSs can connect people and can be used to share information in real time. SNSs such as Facebook, Twitter, and RenRen are becoming the most influential mediums for building social relations, as well as for the sharing and propagation of information. According to recent announcement, Twitter, one of the largest and the most popular SNSs, passed 255m monthly active users and expects 80% of advertising revenue from mobile users.

After repeated explosive growth in the user population, matured SNSs such as Facebook and Twitter become a necessings in modern life in developed countries. In addition, relatively new SNSs such as RenRen and Sina Weibo, targeted for specific country or language speakers, replicate the eruptive expansion of the earlier SNSs. For example, an influential user can be exploited by a person working in online marketing to maximize the marketing effect; malicious users (attackers) disseminate false information or fraudulent messages for the purpose of phishing, scam, or malware intrusion. That is, the

attackers post multiple unrelated messages with trending topics to attract normal users and encourage them to click the malicious links in the messages.

Spam refers to unwanted messages from unknown sources (attackers). One of the major negative aspects of SNS is spam. In the early Internet era, spam appeared in emails or SMS (short message service). However, the domain of spam expanded into SNS as the popularity and usage of the services continued to increase. False information from SNS can spread rapidly in real time. Follow spam was reported recently and is a system that tries to increase the number of relations (or friendships) in users' networks for the purpose of sending spam via SNS. The attack pattern of the follow spam begins with the attacker disseminating spammer accounts that follow a large number of normal users for the purpose of receiving a follow-back or drawing attention to the spam account [27]. Due to the consequent exposure of the public to spam content, this practice definitely lowers the reliability of SNS.

In practice, Twitter has experienced Follow spam problems, reducing users' trust in message distribution and increasing computation overhead. In 2008, Twitter officially announced that Follow spam accounts had followed so many people that they threatened the performance of the entire system. Even with the emerging threat from Follow spam, it has been barely investigated or researched. A contents-based spam filtering approach is employed in the Twitter spam field [21, 9, 45, 64]. However, since spam contents keep changing to avoid content-based detection by inserting URLs and images in spam messages, the contents-based approach is vulnerable against evolving message patterns. To overcome the limitations of the content-based approach, a new approach

using inherent properties of ONS was introduced.

[27] first emphasized that Follow spam should be detected by using its link farming property. They proposed a PageRank-based ranking algorithm to lower the impact of spammers. However, this approach can be burdensome since it needs to utilize social network data for the entire network (i.e., all information for nodes and edges). Therefore, it has a high computational cost and can barely detect Follow spammers in real time. As a result, a novel detection mechanism with low computational cost and real time spam filtering is needed while maintaining the detection performance. In this paper, I suggest two social network-based detection schemes for countering Twitter spam. First, spammer accounts are filtered out with the use of a Triad Significance Profile (TSP) that measures the structural differences between the frequencies of 13 isomorphic subgraphs. I discovered that TSP of a spammer account is different from that of a normal user's account with only 1-hop social networks. According to my experiment, 92.1% of spammers are classified correctly when I used only TSP features for classification. This result suggests that frequency and distribution of isomorphic subgraphs could be informative features for identifying spammers. Secondly, I expand status theory to 'hierarchical homophily' by applying hierarchical gap. My experiments on real Twitter datasets clearly show that my three mechanisms, TSP-Filtering, HH-Filtering, and Hybrid approach are very practical for the following reasons. First, my approaches require only a small user-related 1-hop neighborhood social network called 'ego-network'. Actually, there are only few existing works focused on small neighborhood graph in other areas [48, 2], but none of them discovered the power of neighborhood social network clearly. Therefore, they can be applied to spam

detection systems in social networks as real time solutions. Second, service providers can maintain the credibility and reliability of their SNSs by applying my approaches. Normal users are less likely to be blocked by the system with low false positives (0.01%). Also, a high proportion of true positives (99.4%) provides a secure environment for users. Moreover, I provide a novel spammer detection approach based on structural analysis and relational semantic analysis in ego-network.

The main contribution of this paper is summarized as follows:

- For the first time, I discovered the feasibility of structural information of social network such as triad frequency for classifying spammers and normal users on Twitter;

- I discovered the existence of homophily in terms of social hierarchy. A user's influence on society defines the social hierarchy. In OSN, the social impact could be interpreted as information propagation power. Also, I found that spammers have less hierarchical homophily than normal users by quantitative measurement. I estimated the status gap, hierarchical gap by status binning, and assortativity to find insights. Also, this feature can differentiate spammers from normal users as a classification feature. I conducted a spammer detection experiment in real world Twitter and Weibo datasets;

- My approaches involve more lightweight computation for real time spammer detection than the previous scheme (i.e., global information). Since to check whether a certain user is spammer or not, I only focused

on the ego-networks of each user (i.e., local information);

- To the best of my knowledge, my approaches are the first experiments with real world data to provide credible and reliable Twitter and Weibo system with true positive results of up to 99.4%. I believe that my findings can provide valuable insights into the area of spam detection and defense in various social networks;

Chapter 2

Related Work

2.1 OSN Spammer Detection Approaches

2.1.1 Contents-based Approach

Twitter contents such as user profiles, tweets, and the activity log provide various options for distinguishing spammers from normal users. Spammers generally write tweets that contain a hashtag and URL according to the following research studies that analyzed commonly used hashtags and URL: [21, 9, 45, 64]. COMPA [21] detected compromised accounts that wrote spam tweets based on the tweeting language of the user's account, the tweeting time window, the URL, and the "mention" receiver. This is a personalized detection approach that learns the previous behavioral pattern of each user. Benevenuto et al. [9] and Martinez-Romo et al. [45] proposed classification models that learned the number of hashtags and URLs [9] or spam URLs that are used in spam groundtruth tweets. Yardi et al. [64] studied spammers' strategic behavioral patterns and also concluded that the use of hashtags related to trending topics is a very effective spamming strategy. Gao et al. [26] built a template based on the sentence structure of spam groundtruth tweets and used template matching to filter out spam tweets.

2.1.2 Social Network-based Approach

Network-based spam filtering is based on graphical features of social networks. Node importance estimation algorithms, such as PageRank [49] and HITS [40], or variations, are often employed for spam detection. They have been used extensively in the detection of Web spam [31], which is well suited to detecting similar SNS attacks. Researches on Web spam detection mainly used link based features. Firstly, [3] used the number of inlinks, outlinks, and outlinks per inlink ratio. The author said that search engines or corporate sites (e.g. influential or important pages) are usually very low in outlinks per inlink ratio. [7] and [6] commonly used degree, PageRank, and TrustRank score. But [6] added revised PageRank with a modified damping factor. [20] utilizes the clustering coefficient of a page. NFS (Network Footprint Score) [65] is a spammer detection approach that captures social campaigners by quantifying the likelihood of spam campaign targets based on their PageRank scores. Jiang [37] computes spammers' synchronicity through HITS-based analysis and use the synchronicity to detect fake followers of specific Twitter accounts. Similarly, Viswanath [54] used PCA-based behavioral analysis to detect accounts that increase the popularity of certain pages on Facebook. Ghosh [27] and Boshmaf [12] adopted a random walk-based ranking algorithm. In particular, Ghosh [27] detected the spam linking in Twitter using CollusionRank algorithm. Boshmaf [12] devised a scalable solution that effectively improved SybilRank [13], a ranking-based spam detection method. Most ranking algorithms require global graph information which may not be obtained easily. Some researches in detecting anomalies in OSNs interpreted social networks as heterogeneous network [23, 30] and similarity-based network [68, 70]. They used synthetic

social networks made of the inherent relationship between entities.

2.1.3 Subnetwork-based Approach

There are several network-based schemes that only use local networks called ego-networks [25]. These methods [2, 48, 36, 55] are based on the fact that ordinary users and spammers have different motif occurrences at the ego-network level. However, many OSN spam detection approaches use network-based features additionally to contents-based or behavior-based features. The authors of [55] directly crawled Twitter's data and analyzed them with both contents and social graph modeling-based approaches. Based on the analysis of the contents, categorized into legitimates and spams, they proved that their proposed reputation feature has the best performance among all social graph-based features for detecting abnormal behaviors. However, they only considered the relationship between outdegrees and indegrees in a simple Twitter graph for the proposed reputation feature. Even though this scheme also utilizes a small graph (subgraph), a sophisticated graph design is only part of the triad approach. The authors of [48] used neighborhood subnetwork (i.e., ego-network) to detect comment spammers on Youtube. They also utilized selected discriminating motifs and analyzed them in Youtube video-user relation network. It seems very similar to my work, but it used spam campaign-related motifs. Therefore, it cannot distinguish spammers when they use other sophisticated strategies. [2] extracted weighted subgraphs from the target network and utilizes them as discriminating features to detect spammers. It also analyzed subgraphs by types of anomalies. Based on power-law characteristic of the social network, it compared spammer to normal user's neighborhood

subnetwork in terms of edge or weight distribution.

2.1.4 Behavior-based Approach

Behavior-based spam approaches identify spammers based on the difference in the daily activities (language, usage time, location, friend selection criteria, etc.) between spammers and normal users. Li [44] modeled the user's behavior appeared in web page click sessions to detect the click spam in a search system. They classified cheating sessions using Average Markovian Likelihood. COMPA [21] proposed a mechanism to detect compromise attacks through account hijacking using Twitter usage patterns like language and activation time. Tian [52] also proposed a crowd fraud detection scheme to detect click spam, which also detected spam based on traffic moderateness, target synchronicity, and temporal synchronicity from the sequence of web click actions. SynchroTrap [14] uses tuple to represent time-stamped user actions to detect malicious account groups that generate 'like spam' on Facebook or 'follow spam' on Instagram. They identified colluding groups by clustering groups based on the synchronicity of action tuples. VOLTIME [21] models inter-arrival time patterns in terms of writing reviews, and it detects anomaly users who show the patterns far away from that distribution. Zheng [70] utilized the campaign time window to detect spam campaigners who reside in User-Review sites. So, the behavior-based approach needs to analyze various kinds of data like contents and activity logs.

2.2 Link Spam Detection

Link spam has been widely studied in the web spam detection field. This type of spam is presented as numerous links from a large number of web pages to a few target web pages. Studies on Link spam have been receiving attention due to the limitations of PageRank [49] and HITS [39]. Thanks to significant link characteristics, many weblink graph structure-based spam detection approaches have been introduced [31, 59, 41, 7, 60, 16]. TrustRank [31] is one of the most popular Link spam detection algorithms. It propagates the 'non-spam' label through social networks. Likewise, BadRank [59] propagates the 'spam' label through social networks. Compared to PageRank [49], these two algorithms utilize 'non-spam' and 'spam' label propagation to lower the rank of spam webpages. [8] proposed an advanced Link spam detection algorithm using both 'spam' and 'non-spam' label propagation. These label propagation algorithms require seed knowledge such as a set of spam nodes and a set of non-spam nodes. Therefore, noise in the initial dataset can be a critical issue for these algorithms.

2.3 Data mining schemes for Spammer Detection

In the spam detection problem, most of the existing studies related the problem to the classification task as follows. In general, spam classifiers firstly learn features extracted from SNS using patterns of normal users such as the number of followees/followers, post uploading time and contents information

of user profiles and posts. Then, the classifier determines if a newly given test user is a spammer or normal user by comparing it to the learned pattern. Therefore, if the test user's behavioral pattern feature is far from the normal user's pattern feature (learned feature), the classifier could classify and detect the user as a spammer. In some cases, classifiers adopt a classification threshold to handle the tradeoff between true positive and false positive. Since reliability and credibility are crucial in using SNS, low false positive is treated particularly according to the spam detection system.

In detail, [38] used linear regression for classifying and detecting spammers and it stated that deviant users from normal users' patterns could be classified as spammers. Similarly, [54] utilized PCA (Principal Component Analysis) and it detected Facebook spammers who are distant from the principal component of normal users. Also, Markov random field-based spam classification approach was proposed in [22]. Especially, contents-based spam detection approaches largely used Naive bayes classifier or SVM classifier with contents-related features. In the early stage of spam detection, [29, 35] and many similar studies analyzed token or word in spam contents and applied extracted features to the Naive bayes classifier. [50] proposed an optimized version of SVM spam classifier and achieved efficiency than previous ones. [71] relieved false positive problem by adopting a boundary region to classification result. Since most of the spam classification is the binary classification of spam and non-

spam, ternary classification gives three classification labels including boundary region which means reconsidering region.

2.4 Sybil Detection

Most SNS spam detection systems rely on Sybil detection algorithms. Peer-to-peer systems consist of multiple nodes with several connections (edges). The system has to ensure that each node is clearly identified; otherwise, a malicious user (Sybil) can attempt to create multiple fake identities masquerading as honest nodes [19]. They can then manipulate the system (by zombie machines) or attack the system in order to gain illegal profit such as positive feedback in the reputation system, getting more votes in internet polls, or targeting sites to increase their rank in Google PageRank. There are two main approaches to the Sybil attack: centralized and decentralized. Centralized defense obtains admission control through a central authority. Decentralized defense has no trusted central authority and controls the IP address by binding an identity. For the decentralized attack, SybilGuard [67] proposes that when each node receives $\sqrt{k}$ independent samples from a set of honest nodes of size k, a random walk can be performed to try to discover the Sybil identities by using the intersection probability between honest and Sybil groups. SybilLimit [66] is an enhanced method introduced by [67]. They reduced the attack edge bound in near optimal by exploiting various random walk methods. GateKeeper [53] adapts the ticket distribution algorithm to obtain each node's probability of Sybil/honest users.

Secondly, the centralized method. SybilInfer [18] assumed that the central authority knows the entire social network. After random walks, each node is assigned a Sybil/honest probability by measuring the Bayesian inference. SybilDefender [58] assumed that when starting a random walk in Sybil nodes, it will pass the intersection between honest and Sybil nodes. These approaches apply community detection algorithms to find Sybil communities. SumUp [53] addresses the vote aggregation problem by considering each voter's trust graph and calculating a set of max-flow paths from all voters.

Currently, there are many Sybil detecting methods with various social network properties. SybilRank [13] investigates each node by assuming that honest nodes will have higher degree-normalized landing probability. A random walk is performed to measure the ranking to determine whether the account is Sybil or not. SybilShield [51] utilizes a multi-community social network structure environment, considering sociological properties to cut the edge between honest and Sybil groups, performing modified random walks and figuring out the properties of multi-hop edges. SybilBelief [28] detects Sybil nodes based on a semi-supervised learning framework. This method modifies the Loopy Belief propagation system and the pairwise Markov random field to define each node's classification (Sybil/honest).

Chapter 3

Triad Significance Profile Analysis

3.1 Motivation

Like Web, where the importance of each page is largely determined by who references whom, the influence of individuals on many SNSs is determined by the number of indexes they receive. For example, the number of followers is the most important factor on Twitter and determines social capital, while the number of “likes” on Facebook is similar. This feature, however, has attracted a plethora of frauds who try to increase the importance or reputation of entities by generating bogus indexes, leading to the definition of the spamdexing class of attacks. Twitter’s size has expanded exponentially over the past several years and it now has over 255 million active users after a succession of rapid growth spurts that resulted in an average annual growth rate of 25%. Notably, the social-interaction structure of Twitter is very interesting. Users can follow famous persons–usually celebrities or standout opinion leaders–that they are unacquainted with, as well as close friends. Therefore, Twitter plays an information-propagation role in addition to the role of an online social network [42].

More importantly, contrary to the Weibo, Facebook, and many other social

networks where spam indexes usually originate from fake accounts and circle of colluding link farms, a malicious person can collect followers or fans from innocent, social-capital-conscious users on Twitter. Further, the high rate of follow-backs makes the detection of Twitter spammers more difficult because they receive many normal followers just by following target users [27]. Existing link-farm-detection methods well fitted for web spam detection field, therefore, lose much of their effectiveness in the detection of spamdexing on Twitter.

In this paper, I demonstrate the feasibility of a cascaded SNS-based security scheme to detect Follow spam. Different from the unpractical and heuristic approaches of previous works, with the characteristics of follow-backs I apply triad frequencies and status theory for the first time in my Follow spam detection scheme. Note that the main purpose of this study is not the attainment of engineering optimization for the performance enhancement of prior schemes, but rather, it is the examination of the feasibility of a social-network-based security scheme in a popular online social networking site, i.e. Twitter.

Before I formalize the problem, I address the characteristics of Twitter. All 13 types of directed social graph models and social status with local information can be observed on Twitter. Additionally, Twitter has well-defined social relations in the form of the “follower” and “friend” relationships. In addition to these characteristics, spams show up frequently on Twitter. I practically exploit the policy of Twitter against spams to design my proposed scheme. Twitter’s spam policy is summarized as follows:

- “If you have a small number of followers compared to the number of people you are following”, the account may be considered a spam

account.

- “Multiple duplicate updates on one account” is a factor used to detect spam.

- “If your updates consist mainly of links, and not personal updates”, it is considered spam.

The first policy is related with the social-interaction structure of Twitter while second and third policies have to do with spam contents. Most previous works focused on contents analysis or full information usage of social networks with a high amount of computational overhead. Different from previous approaches considering second and third policies, I accurately detect Follow spam using only local information of the social-interaction structure of Twitter. That is, my cascaded social network scheme is applicable regardless of the content such as Tweet, time and links.

The concept of link farming originated from Web spam. The intent of link and Follow spam is to increase the population of a specific (target) website or reputation. Since normal search engines (e.g., Google) place popular websites on the first page, link-farming websites create numerous links to the target website.

PageRank [49], the most popular website ranking algorithm, ranks websites based on the indegree of the site. Actually, the popularity of inlink nodes is also important, but numerous inlinks are likely to increase the target website’s ranking. Therefore, link farms generally contain plural links, and the links are

created from many nodes to a few target nodes.

Follow spam, a special attack strategy on Twitter has been shown to be a link farming technique. Figure 3.1.1 shows an example of Follow spam.

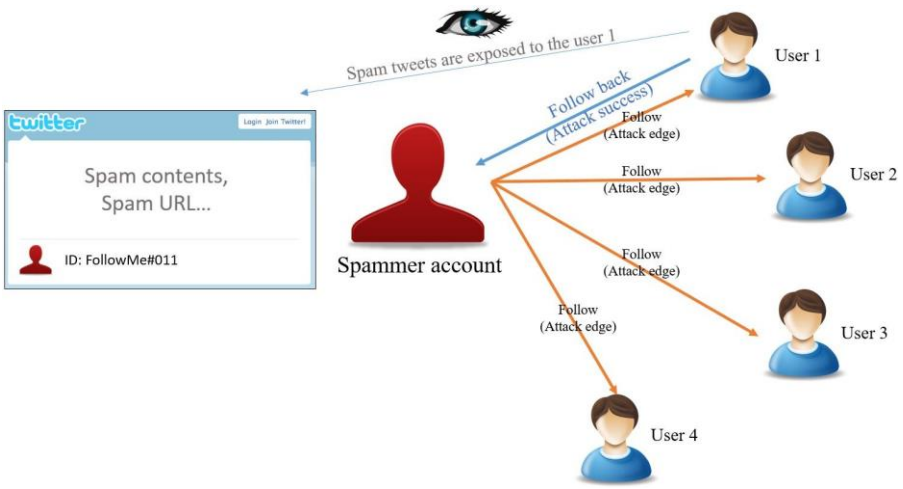


Figure 3.1.1: Overview of Follow spam

Follow spam consists of numerous links, but some differences exist. First, links are created by a few spammer nodes and they target many normal user nodes. More specifically, original link spam denotes many spammer nodes-few normal nodes relationship while Follow spam denotes a few spammer nodes-many normal nodes. Second, the purpose of Follow spam is not just linking, but receiving a follow-back (reciprocal link). A user on Twitter can see tweets (contents) from another user when he/she follows (subscribe) the other's account. Consequently, spammers need to be followed by other users to show their spamming contents such as URL, image and advertisement.

Therefore, to gain more followers and attention, spammers send a large number of following links.

links to normal users. Surprisingly, the majority of followers who Follow spam accounts have been previously targeted by spam accounts. To be specific, 82% of normal users send a follow-back to spammers [27]. If s is a spammer account and his/her outlinks are all attack edges for follow back, the attack strength of s ($\mathrm{AS(s)}$) is defined in (1). I defined the ratio between successful follow spam links (follow back links) of spammers ($N_{fb}(s)$) and total follow spam links of s ($N_f(s)$) as $\mathrm{AS(s)}$ as follows :

$$\mathrm{AS(s)} = N_{fb}(s)/N_f(s) \qquad (1)$$

In equation (1), $N_f(s)$ has the same meaning of "outdegree of *s*". Therefore, the attack strength ($\mathrm{AS(s)}$) of follow spam relies on a successful number of follow backs.

3.2 Twitter Dataset

I conducted an experiment with a large-scale Twitter-follow link dataset that was provided by MPI-SWS [17]. This dataset was collected in September 2009, contains 1,963,263,821 directed social links, and the number of corresponding users is 54,981,152. I also used the Follow spammer dataset from [27] that contains 41,352 spammers as the ground truth.

Table 3.2.1: Twitter dataset

The number of total users	The number of spammers
54,981,152	41,352

Table 3.2.2: Performance estimation of Collusionrank [27]

	True Positive	False Positive
Spammer	94.0%	9.9%
Normal user	90.1%	6.0%

Table 3.2.1 shows the Twitter dataset used in my experiment. I compared the performance of the proposed method with that of Collusionrank [27]. Collusionrank lowers the influence scores of users who connect to spammers and filter out those users who gain high rankings by link farming. It is a user-ranking algorithm based on PageRank. Since I used the same dataset as Collusionrank, I compare the performance of the proposed method with the true positive and false positive results of Collusionrank. According to [27], Collusionrank detected 94% of the 41,352 spammers that appeared in the last low ranked scores 10% of ranking positions; consequently, I could extract the false positives of normal users (9.9%) from Collusionrank with a detection threshold of 10%. I reiterate that the detailed performance of Collusionrank is not described in [27], except for the true positives for spammers within the threshold of the last 10%. Table 15 is the estimated performance of Collusionrank from a true positive value of 94%.

Collusionrank has good performance in terms of true positive and false positive, but it has some limitations as follows:

First, it needs to analyze every node and edge in a social network. The PageRank-based algorithm typically estimates every node's reputation or ranking depending on the reputation of other nodes and edge formation. However, to classify spammers, computing ranks on every node is not practical. In real SNSs, spammers disseminate spamming contents simultaneously. Therefore, a real time spam filtering approach is more effective; fast spam filtering significantly decreases the number of victims of spam. As such, analyzing all social network information is not very pragmatic.

Second, it has a high proportion of false positives in detecting normal users. If 9.9% of normal user accounts on Twitter were blocked, most people would stop using Twitter. A high number of true positives in detecting spammers is also crucial; but the credibility and reliability of the service are maintained by keeping the number of false positives low.

In the following sections, I propose cascaded social information-based spam detection mechanisms that overcome the limitations of Collusionrank.

3.3 Indegree and Outdegree of Dataset

Since Follow spam has a link farming property that involves creating many outlinks, I should investigate whether spammers in Twitter have a higher outdegree than normal users. Also, based on Twitter's spam policy, I focus on

the ratio of the indegree to the outdegree for both normal users and spammers.

In this paper, I use randomly selected 1,000 normal users and 1,000 spammers as the experimental dataset. I determined a large enough sample size with a 95% confidence level and 5% confidence interval.

Table 3.3.1 is the average indegree and outdegree of normal users and spammers.

Table 3.3.1: Average indegree and outdegree

	Average indegree	Average outdegree
Spammer	303.6	866.5
Normal user	401.5	462.0

Table 3.3.2: Performance evaluation using only indegree and outdegree

Classifier	Type	True Positive	False Positive
J48	Spammer	83.9%	19.3%
	Normal user	80.7%	16.1%
RandomForest	Spammer	80.8%	19.6%
	Normal user	80.4%	19.2%

Inevitably, spammers tend to have approximately two times as many outdegrees as normal users. The most interesting observation is that the ratio between the average indegree and outdegree shows significant differences between normal users and spammers. The average indegree and outdegree of normal users are similar and the ratio between the two is 0.86. However, the

ratio between the average indegree and outdegree of spammers is 0.35. This indicates that the indegree and outdegree could be roughly informative for classifying spammers. To classify spammers by only indegree and outdegree, I used J48 and RandomForest classifiers built in Weka. Both algorithms are decision tree-based classifiers. While J48 generates only one decision tree, RandomForest corrects overfitting problems by constructing multiple decision trees during the training process. Table 5.2.1 is the classification performance evaluation using only indegree and outdegree.

As mentioned in the Twitter spam policy, I proved that the number of outdegrees can be a highly useful feature for spam classification. However, a comparison using only the number of degree types between Follow spams and normal users is not enough of a performance measure to inspect spammers as shown in Table 3.3.2. To make up for the spam detection issue, I tried to apply TSP and SS as described in the next sections.

3.4 Twitter spammer Detection with TSP

A prior study showed that, interestingly, several types of networks from different fields such as biology and the social sciences share common properties. In particular, [47] showed that some of the 13 isomorphic triad types are overrepresented while some are under-represented. To the best of my knowledge, I first used this fact to discern Twitter Follow spam. In terms of a social graph, a user is a node and a follow from a person to another person is a directed link from the follower (the person) to the followee (another person). Figure 3.4.1 shows the 13 isomorphic triad classes introduced by [56].

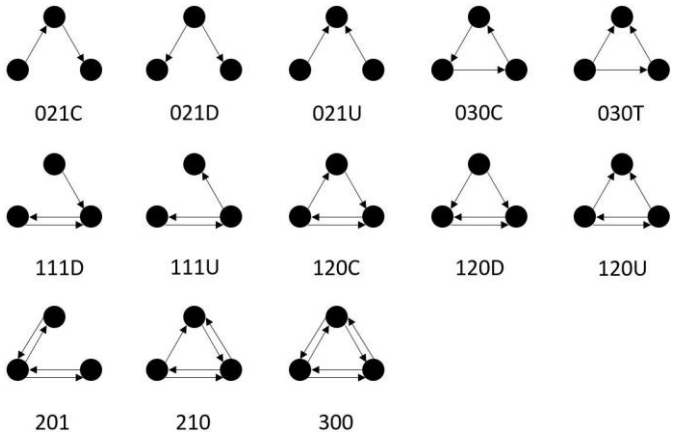


Figure 3.4.1: 13 isomorphic triad classes for analyses

Note that a Follow spammer inevitably generates many follows (directed links) to receive follow-backs (redirected links). For each spammer, I found all of the corresponding triads and counted the frequency of the 13 isomorphic triad classes (for detailed representation of the triad classes, refer to Figure 3.4.1). I performed the same procedures with normal users and compared the differences between the frequency of each triad class for both the spammer-centric triads and the normal user-centric triads. I argue that the triad frequencies of real social networks are different from those of spammers. The triad frequencies of spammers are similar to those of random networks with the same graph properties including the average indegree and the average outdegree.

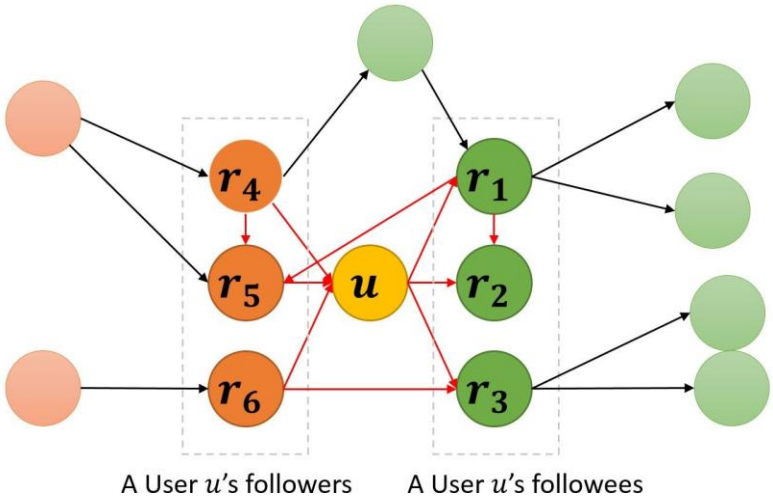


Figure 3.4.2: A user u's ego-network graph G_u (red-colored edges and named nodes)

For a given local network G_u of a user u as shown in Fig. 3.4.2, I estimated the number of occurrences for each triad class. G_u consists of social links between u and 1-hop neighborhoods of u. Suppose that u is following r_1, r_2 and r_3 and is also followed by r_4, r_5 and r_6. In this case, r_1, r_2 and r_3 are "Followees" of u. In the same manner, r_4, r_5 and r_6 are "Followers" of u. Also, there are directed social links between them (represented as red-colored links in Fig. 3.4.2). To determine whether user u is a spammer or not, I analyzed user u's social graph G_u consisting of 7 nodes and 10 edges. This is a subgraph of a Twitter social network, and every user can have his/her own social network.

To discover the phenomenon whereby spammer social network comprise

subgraph features that are different from normal user social networks, I compared spammer triad frequencies with those of normal users. For each triad class *i*, the statistical triad occurrence is described by the Z-score Z_i [47] in Equation (2).

$$Z_i = \frac{N_{spam_i} - <N_{legit_i}>}{std(N_{legit_i})} \quad (2)$$

where $\mathrm{N_{spam_i}}$ is the occurrence number of the triad class i in a spammer's network, and $< \mathrm{N_{legit_i}} >$ and $\mathrm{std(N_{legit_i})}$ are the mean and standard deviations of its appearances in the legitimate user networks, respectively. The TSP is, therefore, the vector of the Z scores that are normalized to length 1 in equation. (3).

$$TSP_i = \frac{Z_i}{(\sum {Z_i}^2)^{1/2}} \quad (3)$$

To visualize this insight from network comparison, I computed the average vector of TSP for 1,000 spammers and normalized it. I also computed N_{legit_i} based on 1,000 legitimate users.

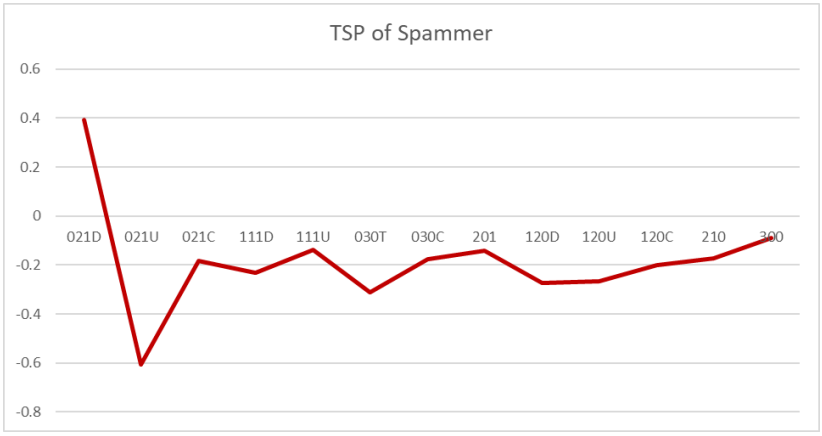


Figure 3.4.3: Average TSP of spammers and normal users (y=0 line). Error bar means the standard deviation of spammers' TSP.

Figure 3.4.3 compares the TSPs of spammers and normal users (y=0 line). Normal users generally have more triads compared to spammers, meaning that the neighbors of normal users are socially well connected with isomorphic triad patterns; therefore, this phenomenon produced more triad counts overall. Alternatively, spammers have lower triad counts than normal users because their 1-hop neighbors are not likely to acquaint themselves with the other 1-hop neighbors.

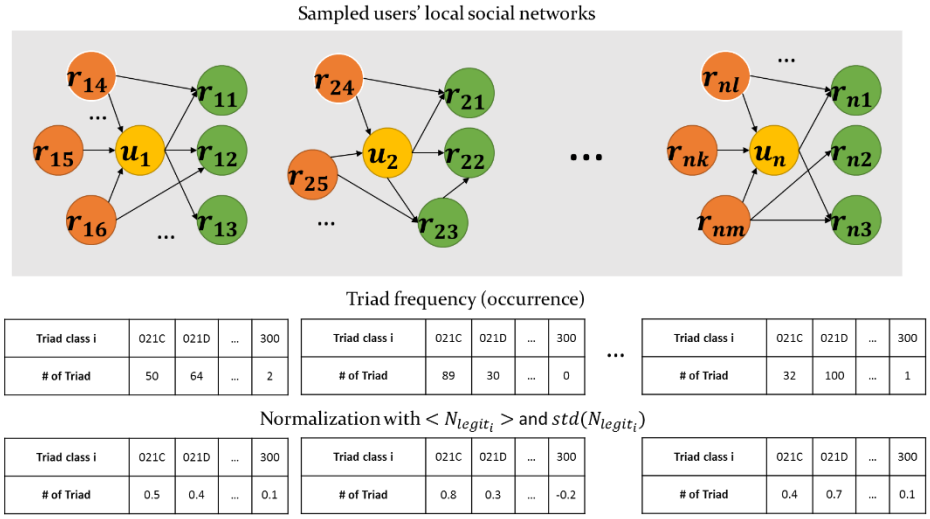
Since spammers usually select their followees randomly, there are few connections between spammers' neighbors. Triad 021D, however, indicates exceptional triad counts, whereby spammers have more 021D triads than normal users. The 021D triad class represents the plural-following actions from a node. It also represents link-farming activity. Since the actions of Follow spammers involve the production of numerous out-links, their high 021D triad counts make sense. The distinction between the TSPs of spammers and normal

users therefore explains why my TSP-detection approach is feasible.

As mentioned earlier, I randomly sampled sets of 1,000 spammers and 1,000 normal users from the original dataset [27] and conducted an experiment with TSP. I determined that the sample size was large enough with 95% confidence level and 5% confidence interval.

3.5 TSP-Filtering

The following process was used for the applicable value of the TSP-Filtering based on the experiment. First, I obtained the mean and standard deviations of each frequency for the triad class across all of the Twitter accounts. Since 1,000 normal users are sufficiently representative to support every Twitter account (confidence level: 95%, confidence interval: 5%), I computed the mean and standard deviations of the 1,000 randomly-sampled normal users. The mean value of the triad class i is $< N_{legit_i} >$ and standard deviation of the triad class i with $< N_{legit_i} >$ is $std(N_{legit_i})$, respectively. Figure 3.5.1 shows the sampled user's local social networks and triad frequency normalization.



Triad class i	021C	021D	...	300
# of Triad	50	64	...	2

Triad class i	021C	021D	...	300
# of Triad	89	30	...	0

Triad class i	021C	021D	...	300
# of Triad	32	100	...	1

Triad class i	021C	021D	...	300
# of Triad	0.5	0.4	...	0.1

Triad class i	021C	021D	...	300
# of Triad	0.8	0.3	...	-0.2

Triad class i	021C	021D	...	300
# of Triad	0.4	0.7	...	0.1

Figure 3.5.1: Triad frequency normalization

Second, I counted the spammer-triad frequencies and the normal user-triad frequencies for every social-network subgraph of every user account; the triad frequency represents the triad appearances in each user network. Then, I normalized the frequencies with $< N_{legit_i} >$ and $std(N_{legit_i})$ (Figure 3.5.1). In the case of the spammer, I can use Equation (2); however, in the normal user's case, I can use the re-translated Equation (4), where N_{spam_i} is the occurrence number of the triad class i in a normal user's network:

$$Z_i = \frac{N_{legit_i} - < N_{legit_i} >}{std(N_{legit_i})} \qquad (4)$$

3.6 Performance Evaluation of TSP-Filtering

I conducted the experiment using J48 and RandomForest implemented in Weka (10-fold validation). Table 3.6.1 shows the performance evaluation results for the TSP method without indegrees and outdegrees. Table 19 shows the performance evaluation results for the TSP method with indegrees and outdegrees. On the other hand, Table 3.6.2 shows the performance evaluation results for the TSP method with indegrees and outdegrees.

Table 3.6.1: Performance evaluation using TSP-Filtering (w/o indegree and outdegree)

Classifier	Type	True Positive	False Positive
J48	Spammer	91.0%	9.4%
	Normal user	90.6%	9.0%
RandomForest	Spammer	92.1%	8.4%
	Normal user	91.6%	7.9%

Table 3.6.2: Performance evaluation using TSP-Filtering (w/ indegree and outdegree)

Classifier	Type	True Positive	False Positive
J48	Spammer	91.7%	9.2%
	Normal user	90.8%	8.3%
RandomForest	Spammer	92.3%	7.6%
	Normal user	92.4%	7.7%

From Table 3.6.1, even without indegrees and outdegrees, TSP-Filtering for RandomForest has a powerful spam-classification performance with 92.1%.

From Table 3.6.2, the proposed approach with indegrees and outdegrees has 92.3% true positives and a lower proportion of false positives (7.6%) than Collusionrank (9.9%). Unlike Collusionrank, which needs to analyze every link to rank every node, my TSP approach is a fast and low-cost detection mechanism that uses only the 1-hop-neighborhood network for each user. Therefore, the TSP approach is a more lightweight and efficient mechanism for detecting follow spammers in real time.

To define a preferred sequence of attributes, I measured the importance of feature attributes based on information gain as shown in Table 3.6.3. In Table 3.6.3, feature attributes listed in descending order of information gain. Information gain can be computed as follows:

$$InformationGain(C,A) = Entropy(C) - Entropy(C|A) \quad (5)$$

In Equation (5), C represents the given class such as spammer and normal user. A is the feature attribute. For example, InformationGain(spammer,021D) refers to the amount of entropy decrease in a spammer class when the feature attribute 021D is provided.

Table 3.6.3: The importance of feature attributes based on information gain (TSP-Filtering)

Feature attributes	Information Gain
021D	0.2867
021U	0.2556
021C	0.2366
111U	0.2267
201	0.1418
030T	0.1408
111D	0.1399
120D	0.136
120U	0.1075
120C	0.0871
300	0.0859
210	0.0794
030C	0.0465

As I showed in the experiment results, 021D is the most significant factor in classifying follow spammers because of its property of two out-edges. Follow spammers tend to have many out-edges to normal users. This tendency is presented naturally in 021D. The following attribute, 021U, is also significant in classifying normal users because of its two in-edges. Normal users are likely to have more followers than spammers at stable points of the Twitter SNS system. Twitter is a very special SNS due to its subscription characteristics. The more informative the users' contents are, the more followers subscribe to the user. Since most spammers upload advertisements or spamming content on their accounts, they have fewer followers than normal users. Understandably, some normal Twitter users try to follow many users at the initial and transition points for subscriptions or other reasons. However, normal users at the steady

point have a larger number of indegrees (i.e., followers) than outdegrees (i.e., friends) due to effective influence or fruitful content because of the psychology of popularity. In addition, the remaining features of TSP are gradually reflected in the distinction between the follow spammers and normal users because of the social interaction.

Chapter 4

Hierarchical Homophily Analysis

4.1 Motivation

As online social networks (OSNs) such as Twitter and Sina Weibo have evolved to a powerful and convenient information sharing platform, they also have attracted an increasing number of malicious users who spread commercial or unlawful content through OSNs. Fraudulent actions in OSNs often are involved with abusive creation of social links. Fraudulent links unduly increase the fame of spammers and also become information pipes through which unsolicited information being disseminated. One type of OSN attack is spam indexing or spamdexing. Spamdexing – an attack originally employed to improve the rank or popularity of webpages – generates a copious number of artificial links to spammers in order to bloat the fame or popularity of the spammers. Spammers then exploit the unjustly gained fame for easy and wide propagation of their message, often to achieve monetary gains. Another type of OSN attack is ”follow spam 5”, attacks customized for Twitter-like OSNs.

Malicious users first need to establish information conveyers through which unwanted advertisements or illegitimate information can be freely disseminated to victims. One method to establish information pipes is to generate plentiful of follows to randomly selected innocent targets anticipating reciprocal follow backs from the victims. According to [27], people, who receive the following link from spammers, respond with reciprocal follows with the probability of as high as 82%.

I use the term “spam linking” that includes spamdexing and follow spam. Spam linking often will develop into the wide dissemination of abusive contents and incur annoyance and inconvenience to users. Circulation of abusive messages eventually hurts the credibility of the whole OSNs as an information-sharing platform. Detecting and eviction of spammers who create random connections are important in maintaining the healthy online social ecosystems.

In this section, I try to increase the design space of spammer detection adding social network features. Particularly, I propose a novel spammer detection scheme that utilizes the unique social network characteristic called hierarchical homophily. As far as I know, this is the first approach that utilizes the hierarchical homophily property in spammer detection. The reason that I select homophily as a classification criterion is that normal users follow or make

relations with persons whom they choose after some thoughts have stronger homophily that spammers who generate random connections. According to McPherson [46], homophily means that relationships between people who share similar characteristics occur more often than among dissimilar people. For example, people of the same or similar occupation, income, economical wealth, race, education level are more likely to be related than people of dissimilar social characteristics. Recent social network analysis studies [10, 4, 34, 5, 11] confirmed that homophily indeed exists in online social networks such as Twitter reciprocal-reply networks, political conversation logs, sentiment in messages in Twitter, and DBLP co-author network.

However, the major concern is "how to measure homophily in OSNs?". Homophily is a meta characteristic that manifests on top of base properties such as income, wealth, education level and etc. Because most of base properties cannot be observed in OSNs, I resort to adopting social status – a measure that can be estimated by analyzing the connectivity in OSNs as a base property. My preliminary study based on social status confirms the existence of homophily in OSNs. Note that like many other base properties (e.g. income, wealth) of homophily, social status is a hierarchical property that can be quantified. I use a term hierarchical homophily to emphasize the quantifiable property of social status in OSNs. Some previous studies are giving proof of homophily property as a spammer classification feature. [24] classifies social actors such as leaders

(e.g., news groups), lurkers, spammers with link strength prediction based on contextual information. Link strength is estimated with user-user relationship. [62] is a spammer detection method that utilizes group modeling based on network information. Grouping users based on similarity is one of the applications of homophily property.

The two major contributions of my research are as follows. Firstly, I discovered the existence of homophily in terms of social hierarchy. A user's influence on society defines the social hierarchy. In OSN, the social impact could be interpreted as information propagation power. Second, I found that spammers have less hierarchical homophily than normal users by quantitative measurement. I estimated the status gap, hierarchical gap by status binning, and assortativity to find insights. Also, this feature can differentiate spammers from normal users as a classification feature.

I design a novel spammer detection scheme based on hierarchical homophily. I introduce several features that can capture the level of hierarchical homophily of individual OSN users. Note that my method is a completely social network-based approach, and is computationally efficient because it requires only the user's ego-network for the classification. I carried out a performance analysis of the proposed scheme with two real-world datasets obtained from Twitter and Sina Weibo. It is worthwhile to note that while Twitter dataset contains follow

spams, Weibo dataset is mostly concerned with spamdexing. Experimental results show that my proposed method has higher detection rates (97.6% on Twitter and 99.1% on Weibo) and lower false positive rates (2.4% on Twitter and 0.2% on Weibo) than existing methods.

4.2 Hierarchical Homophily in OSN

In this section, I verify whether hierarchical homophily exists in online social networks, such as Twitter and Sina Weibo. In offline social networks, the socioeconomic status of an individual is often determined by her income, wealth or occupation [57]. However, since an online social network is an anonymous society with its own ecosystem, socioeconomic information of subscribers is difficult to obtain or estimate. Therefore, it is desirable to develop the base homophily property that can be solely estimated from the graphical information only. Fortunately, several graph-based algorithms that estimate node status or importance have been proposed. PageRank and HITS probably are the two most important ranking algorithms. I defined them as M 2 and M 3. However, both require global network information. To avoid the overhead of collecting the global information and for the ease of computation, I also used a status estimation method that can compute each individual's status from her ego-network only and defined it as M1 (Equation (6)).

- M1: From the viewpoint of social influence, the number of subscriptions that a user receives is the most intuitive measure for estimating her influence or status. Let indegree(u) and outdegree(u) are the number of u's followers and the number of u's followees, respectively. Then user u's status, M1(u), is determined as follows.

$$\mathrm{M1(u)} = \frac{indegree(u)}{indegree(u)+outdegree(u)} \quad (6)$$

  Note that $0 \leq \mathrm{M(u)} \leq 1$ and larger M(u) means higher status.

- M2: PageRank of a user [49]

- M3: Authority score of a user computed by HITS [40]

According to existing researches [61, 1, 17], follower count was a major factor to determine influential opinion leaders in information cascading network. Similarly, PageRank and HITS Authority score, which assess the importance of nodes, were also used for identifying influential users. I used the three metrics as a representative to confirm that homophily is a feasible feature for spam detection.

4.2.1 Basic Analysis of Datasets

Table 4.2.1.1: Dataset description.

	Twitter	**Weibo**
# of users	54,981,152	11,537,323
# of spammers/customers	41,352	395
#of links	1,963,263,821	38,055,283

I examined the existence of hierarchical homophily in OSNs using the real-world datasets observed from two large OSNs: Twitter [17, 27] and Sina Weibo [69]. Table 4.2.1.1 is the description of the two datasets. Twitter and Weibo dataset commonly consist of users' IDs and follow links only. Every link means a unidirectional follow link. Note that both datasets identify spammers that I can use ground truths in evaluation. It is worthwhile to note that the attack strategies of Twitter and Weibo are not the same. The Twitter spammers I experimented with were spammers who are suspended by Twitter, which may include follow spammer. Similarly, other types like Tag Spam, may also exist. In Twitter, many spammers follow randomly selected innocent users expecting that the victims respond with follow backs to the spammer. Once reciprocal follows are established, the spammer uploads spam contents to her own

‘timeline’, so the contents are exposed to follow-back users. As a result, a spammer creates a link to multiple users, which is a one-to-many broadcasting spam attack. On the other hand, the attack strategy in Weibo is that certain users purchase fake followers in the market to increase their followers and eventually to bloat their fame. Spamdexing is considered as a harmful activity because it arbitrarily manipulates the trust and influence of the users in OSNs. This attack strategy is regarded as a distributed type, where a number of paid users create a follow link to one follower buyer (or market customer).

Table 4.2.1.2: Comparison of the number of followers and followees between spammers and normal users on Twitter.

	The number of followers	**The number of followees**	**The number of reciprocal links**	**Reciprocal link ratio**
Spammer	211	860	146	0.070
Normal user	568	548	300	0.288

I analyzed the ego-networks of Twitter users and computed their status. Table 4.2.1.2 shows the basic ego-network statistics such as the number of followers and followees, the number of reciprocal links and the ratio of the reciprocal

links in Twitter. Note that because spammers usually secure good numbers of followers and followees to maximize their influence, I exclude users with less than 10 relations in this study. In Table 4.2.1.2, I can observe that spammers follow 860 persons on average while they receive 211 follows. On the contrary, the numbers of followers and followees are well balanced in the case of normal users. Also, note that the number of reciprocal links of normal users is about twice greater than that of spammers. Similarly, the reciprocal link ratio the ratio of the number of reciprocal links to the total degree of normal users is more than four times greater than that of spammers. Because spammers and normal users have very different statistics, it may be tempting to devise detection schemes based on the basic statistics. However, as shown in later, hierarchical homophily provides a better foundation than the basic statistics for the development of spam detection schemes.

Table 4.2.1.3: Social status of spammers and normal users on Twitter.

Label	Follower (inlink)		Followee (outlink)	
	Average	Standard deviation	Average	Standard deviation
Spammer	0.44	0.14	0.59	0.23
Normal user	0.38	0.20	0.39	0.20

Table 4.2.1.3 shows the social status statistics of spammers and normal users on Twitter. Since M2 and M3 status value of Twitter users are too low, I use M1 social status for this analysis. Note that on Twitter, while a spammer can freely follow randomly selected users, but it is difficult to fabricate followees. I can observe three interesting facts in Table 4.2.1.3. First, both followers' and followees' status of spammers are greater than those of users. Second, the average status of spammer's followees is significantly greater than that of normal user's followees while the average status of normal users' followers and followees are almost the same. Because spammers tend to follow users of many followers, it is not surprising that the status of spammers' followees is large. However, it can be surprising that spammers receive follows from users of higher status than normal users. I guess that follow spam attacks is quite successful in inducing follow backs. Users of high status tend to be highly active users who are conscious to maintain their popularity. Therefore, active users more prone to follow backs to unknown followers than inactive users. Lastly, I can observe in Table 3 that the standard deviation (SD) of spammer's followees is larger than that of spammer's followers. This observation indicates that spammers select targets randomly but they receive follows from rather homogeneous groups of people.

Table 4.2.1.4: Reciprocal neighborhood social status distribution of spammer and normal users.

Label	Average reciprocal neighborhood status	Standard deviation of reciprocal neighborhood status
Spammer	0.41	0.10
Normal user	0.50	0.12

I expect that normal users generally establish reciprocal relations with people whom they know personally [63]. Therefore, reciprocal links may provide better clues for the characterizing of users than one-way relationships. I repeated the previous analysis with reciprocal links only. In Table 4.2.1.3, the followees of spammers have larger social status than those of the normal users. However, in the reciprocal link case (Table 4.2.1.4), the neighbors (followees as well as followers) of the normal users have higher average status. I can conclude that normal users tend to form reciprocal relations with persons with higher status than spammers. Table 4.2.1.5 shows the average and standard deviation of neighborhood status by the ego's status level. This indicates that the status level of normal users and neighborhoods have correlations while spammer does not have a correlation with the neighborhood.

Table 4.2.1.5: Neighborhood social status distribution of spammer and normal user by status level

	Status level	Average of Neighborhood status	Standard deviation of neighborhood status
Normal user	**Low**	0.43	0.12
	Medium	0.49	0.12
	High	0.56	0.13
Spammer	**Low**	0.41	0.10
	Medium	0.41	0.09
	High	0.46	0.11

4.2.2 Status gap distribution and Assortativity

Hierarchical homophily means that two end nodes on a link have the similar social status (small gap in social status). As the first step to investigate hierarchical homophily, I compute the social status gaps between two directly

connected nodes. Figure 4.2.2.1 is the probabilities of status gaps of normal users and spammers, respectively. In Figure 4.2.2.1, I can observe that about 50% of normal users' links have gaps in the interval of [0.1, 0.2]. Surprisingly, status gaps of spammers' links are concentrated on the low range of [0.0, 0.1]. However, status gaps of spammers' links are more widely distributed than those of normal users' links and more than 40% of links have gaps more than 0.2.

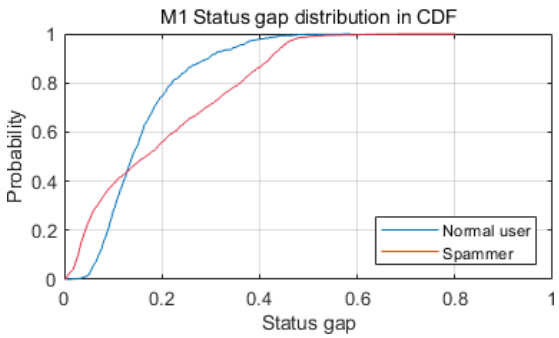


Figure 4.2.2.1: Probabilities of the status gap of normal users and spammers (Twitter).

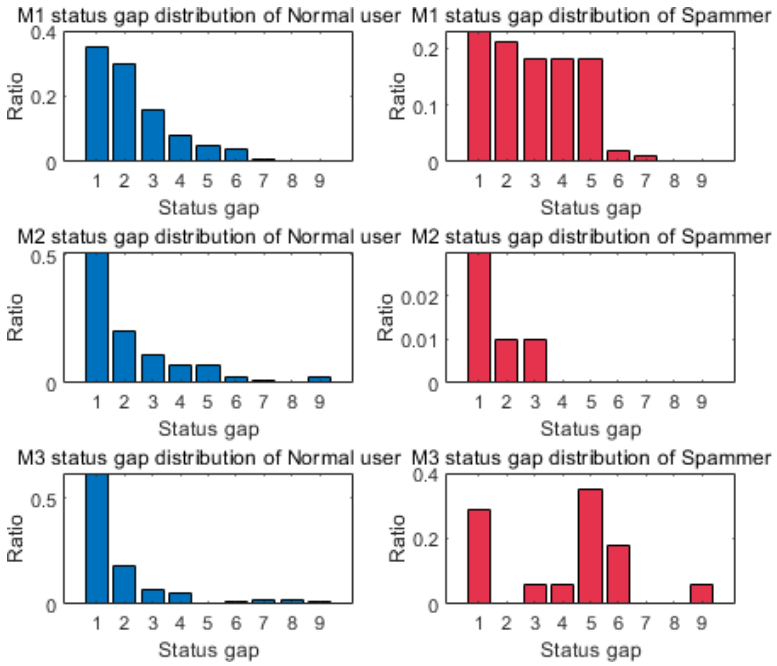


Figure 4.2.2.2: Quantized status gap distribution of normal users and spammers.

In Fig. 4.2.2.1, the probability of the low gap of spammers is higher than that of normal users such that the homophily of spammers can be higher than that of normal users. Probabilities on M2 and M3 cannot show significant difference between normal user and spammer. I conducted further investigation by converting the levels of homophily to quantifiable numbers. I quantize the status gaps into 10 levels. The whole status gap range are partitioned into 10 equal length intervals such that interval k includes status gaps in the range of [k−1 , k]. Fig. 4.2.2.2 shows quantized probability of status gaps. In Figure 4.2.2.2, I can observe that in the case of normal users, the probability decreases rapidly as the status gap increases. On the contrary, spammers have a fairly large distribution in intervals of large status gap. In this respect, the homophily characteristic is stronger for the normal user.

Assortativity is the measure that can quantify the level of homophily. According to [11], the neighborhood assortativity is the possibility that each individual is influenced by the overall social status of all of the people it interacts with. Applying assortativity, I can see the homophily characteristic of normal users and spammers observing what type of correlation there is between ego and its neighbors in terms of social status. In equation (7), let U be the set of all user nodes in the graph, and n be the number of links in the graph. Also, $S(U)$ is the average status of all users, and $\overline{S(K_u)}$ and $\overline{S(K)}$ denote the average status of user u's neighborhood set, and the average status of all user's neighborhood sets, respectively. Let $\sigma(U)$ be the standard deviations of the status of the entire population. Similarly, let $\sigma(\overline{K})$ be the standard deviations of the average status of every ego-networks (Note there are $|U|$ ego-networks).

The Neighborhood assortativity of graph G, A(G), is calculated as follows.

$$A(G) \equiv \frac{1}{n-1}\sum_u \left[\left(\frac{S(u)-<S(U)>}{\sigma(U)}\right)\left(\frac{\overline{S(K_u)}-<\overline{S(K)}>}{\sigma(\overline{K})}\right)\right] \qquad (7)$$

Table 4.2.2.1: Comparison of neighborhood assortativity between normal users and spammers.

	M1	M2	M3
Spammer	0.090	0.195	0.164
Normal user	0.133	0.017	0.030

Neighborhood assortativity has a value between (-1 and 1), and high

assortativity implies a large degree of homophily. If there is no homophily in a graph, then its assortativity will be 0, and negative assortativity means people of low status have prone to make relationships with people of high status and vice versa. Table 4.2.2.1 shows the assortativity of normal users and spammers. Both normal users and spammers have homophily characteristics. However, the assortativity of normal users is larger than that of spammers and normal users' homophily is stronger than spammers' homophily.

Figure 4.2.2.3 visualizes the homophily of normal users and spammers. For each user, either a spammer or a normal user, Figure 4.2.2.3 plots points each of whose x-axis value is its social status and y-axis value is the average social status of neighbors. I can observe that, for normal users, the average neighborhood status increases in proportion to egostatus. But not for spammers, average neighborhood status is almost stationary.

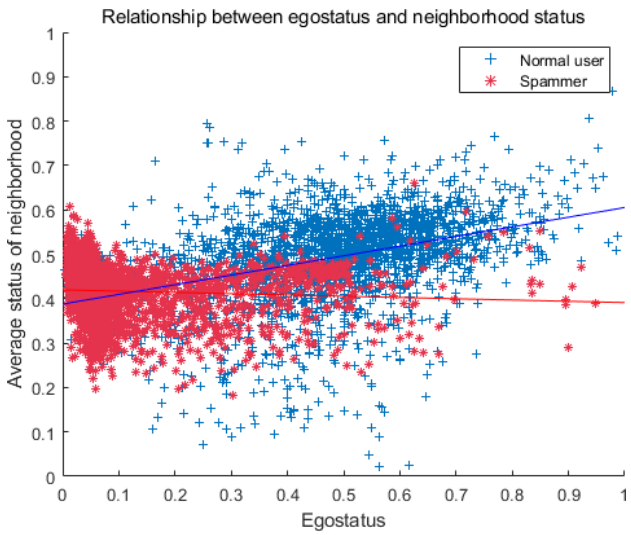


Figure 4.2.2.3: Relationship between Egostatus and average neighborhood

status.

4.2.3 Hierarchical gap distribution

In this subsection, I apply the concept of social hierarchy and analyze how ego is related to higher or lower social status neighbors than itself. To determine whether the users in an online social network follow hierarchical homophily, I investigated the relationship between the hierarchical gap and the ratio of being linked. To estimate the hierarchical gap, I divide 0-1 normalized social status into *N* social classes (hierarchies).

To estimate the hierarchical gap, I divide 0-1 normalized social status into N social classes (hierarchies). I divide entire user into N quantiles such that a hierarchy H_N contains the user nodes whose social status belong to the highest $\frac{100}{N}$ percentile and hierarchy H_1 contains users of the lowest quantile. Then I define the status gap between user *u* and *v*, $\mathrm{G(v,u)}$, as

$$\mathrm{G(v,u)} = H(v) - H(u), \tag{8}$$

where *H(u)* is the quantized hierarchy class that the user *u* belongs to.

The average gap between ego and neighbors ($\overline{G(u)}$) is calculated as follows.

$$\overline{G(u)} = \frac{1}{N}\sum_{v \in N(u)} \mathrm{G(v,u)} \tag{9}$$

where *N(u)* is the set of neighbor nodes of *u*.

Using equation (9), I can determine how much the ego is associated with neighbors with some social hierarchy gap. In equation (8), $G(v,u)$ is a positive value when the status of v, one of u's neighbors, is higher than the status of u. Otherwise, $G(v,u)$ is a negative value.

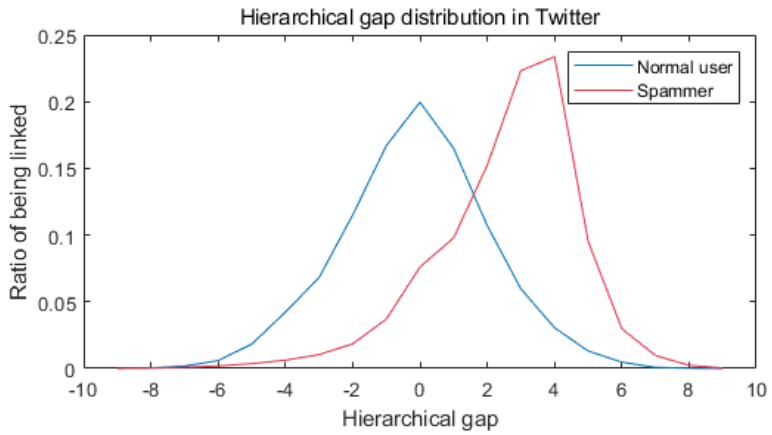


Figure 4.2.3.1: Ratio of being linked by the hierarchical gap in Twitter expressed in positive and negative relationships.

I compute $G(v,u)$ with the Twitter dataset and show its distributions for normal users and spammers in Figure 4.2.3.1 based on M1 social status. Spammers in Twitter normally choose a 'broadcasting attack' strategy, in which a spammer spreads numerous links to unspecified individuals. Figure

4.2.3.1 shows that normal users in Twitter make relationships with users whose social status is similar. I can observe that the probability of being linked is highest when the gap is 0. This means that the normal user and most of his or her neighbors belong to the same social status group. This ratio is gradually decreased as the hierarchical gap increases. On the other hand, the highest point of the spammer occurs in gap 4. The underlying cause of this phenomenon is that most of the followers of a spammer are very active users in making follower/followees. In conclusion, Fig. 4.2.3.1 indicates that in terms of online social status, normal users in online social networks have more hierarchical homophily than spammers.

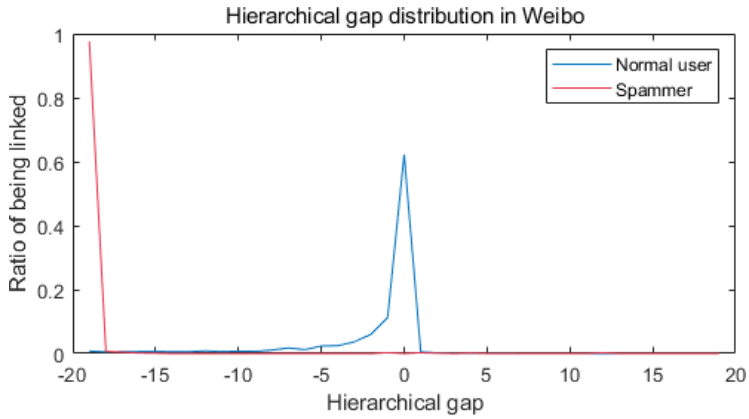


Figure 4.2.3.2: Ratio of being linked by the hierarchical gap in Weibo expressed in positive and negative relationships.

I performed the same analysis with the Weibo dataset and the results are shown in Figure 4.2.3.2. In the case of Weibo, the attackers, the customers who bought

fake followers in large quantities, are the recipients of the distributed attacks. Therefore, I observed only the followers of each user, i.e., neighbors connected by the incoming link. In this case, the normal user receives ‘follow (incoming link)’ from users with a similar hierarchy, while spammers receive ‘follow’ from users with a status/hierarchy much lower than him or herself. These results suggest that most followers of customers may be fake followers with very little social activity. Note that I used N=20 in analyzing the Weibo dataset to emphasize this phenomenon.

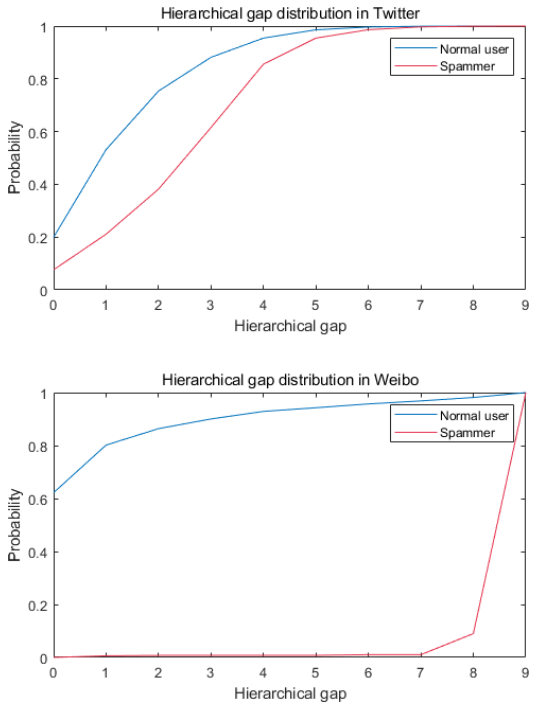


Figure 4.2.3.3: Hierarchical status gap in CDF.

Finally, Figure 4.2.3.3 compares the distributions of G(v,u) for normal users and spammers. I discovered that normal users in Twitter and Weibo have naturally stronger homophily that decreases by increasing the hierarchical gap.

4.3 Performance Evaluation of HH-Filtering

My preliminary analysis described in the previous section revealed that spammers have weaker levels of homophily than normal users. I developed classification features that best can distinguish spammers from normal users. Table 4.3.1 is an abbreviation of the spammer classification features that I devised. F0 is a basic feature that represents the status or importance of each individual. Some prior methods use egostatus for classification. Homophily is a context feature and feature F1 is inevitable in measuring the level of homophily. As shown in the previous section, neighbors of a normal user are relatively more homogeneous than those of a spammer. I introduced F2 anticipating that it is effective in detecting users with heterogeneous neighbors. Finally, feature F3 is similar to the assortativity of an individual.

Table 4.3.1: Features for spammer classification experiment

Index	Features
F0	Social status of ego (a.k.a. egostatus)
F1	Average social status of neighbors
F2	Standard deviation of neighbors' social status
F3	Z-score vector that represents discrepancy at every status gap

Feature F0-F2 are straightforward. However, F3 - the Z-score vector – requires further explanation. Note that I discretize status gaps into 10 levels. For a user u, I define a vector X_u which contains the probabilities that the *u*'s status gap belongs to quantized levels. Similarly, I also define a probability vector y. The vector y contains the average probabilities of normal users. It is worthwhile to note that I obtain the vector y from labeled normal user data from the training dataset. To compute Z-score, I compute the Standard Deviation of normal users' probabilities at each quantile. Assume that a vector D contains the SDs. The Z-score for an interval i is computed as follows (equation (10)):

$$\boldsymbol{Z_u}[i] = \frac{\boldsymbol{X_u}[i] - \boldsymbol{Y}[i]}{\boldsymbol{D}[i]} \qquad (10)$$

Figure 4.3.1 shows examples of Z-scores for the average spammer. Note that the Z-score vector is a $2N$-1 dimensional vector in the case of N quantile.

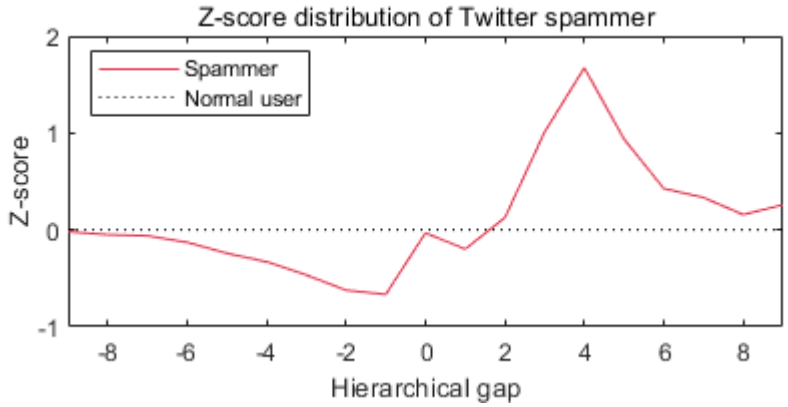


Figure 4.3.1: Z-score distribution of average Twitter spammer.

I used two datasets observed from Twitter and Sina Weibo. I sampled normal users and spammers such that their status distributions are about the same. I discarded samples of less than 10 links. I balanced the numbers of spammers and normal users; in Twitter 1,000 normal users and spammers each is sampled while 364 spammers and normal users each are sampled from the Weibo dataset. In the case of Twitter, I focused on reciprocal relations because the attack strategy in Twitter is “follow spam” that induces follow-backs. However, I used one-directional links in Weibo because the attack type in Weibo is distributed.

I performed 10-fold validation applying various classification schemes. During each validation, 90% of randomly selected normal users and spammers are assigned to the train data and 10% are assigned to test data. I compared

performances of several classifiers including J48, AdaBoost, and Logistic Model Tree and RandomForest. Among them, RandomForest performs best and the results of RandomForest are described. Tables 4.3.2 shows confusion matrices for Twitter and Weibo spammer detection tests.

Table 4.3.2: Confusion matrix of Twitter and Weibo experiment.

Dataset	Label	True Positive	False Positive
Twitter	**Spammer**	97.6%	0.6%
	Normal user	99.4%	2.4%
Weibo	**Customer**	99.1%	0.2%
	Normal user	99.7%	0.8%

From Table 4.3.2, I discovered the hierarchical homophily based scheme have the most dominant discriminating power in spammer classification. Because the proposed method performs almost perfectly in terms of true positives, I paid attention to the false positives in the normal user classification. The false positive is one of the most critical factors in evaluating social networking services. If a spammer detection system falsely classifies a normal user as a spammer, the system bounces off innocent users who loses business opportunities. Therefore, to maintain the reliability of social networking services, a spammer-defense system should endeavor to lower its false positive probability.

Table 4.3.3: Confusion matrix of Twitter and Weibo experiment (egostatus is excluded).

Dataset	Label	True Positive	False Positive
Twitter	**Spammer**	93.6%	5.6%
	Normal user	94.4%	6.4%
Weibo	**Customer**	98.8%	1.4%
	Normal user	98.5%	1.2%

I mentioned that egostatus is a basic feature that many prior methods include for spam detection. I conducted the same tests excluding egostatus in order to investigate the robustness of homophily features as well as to explore its importance in spam detection. In Table 4.3.3, I can observe that the performance of the proposed scheme degrades in terms of both true positive and false negative. Even I deleted the major feature, the true positive is still high in both datasets. The true positive of spammer classification on Twitter is decreased more than that in Weibo. I infer the main reason for this issue is the link-farming property of Twitter spammers. Twitter spammers normally have low social status, because of many follow links to random users. Also, a large number of spammers show colluding actions: they follow each other to increase the number of followers. In conclusion, regardless of egostatus feature exclusion, hierarchical homophily based features are robust on spammer classification.

Chapter 5

Overall Performance Evaluation

I compared the performance of my proposed scheme with those of several prior social attack detection schemes; SybilRank, NFS, CatchSync. I note that every baseline method needs only network-based features like my proposed approach. SybilRank [13] is a graph-based ranking algorithm targeted for search engine optimization attacks. It penalizes Sybil activities and lowers artificially-made influential nodes. CatchSync [37] is a HITS-based user ranking algorithm that locates spammers on lower positions of the ranking. The major intuition of this approach is that anomalies in online social networks tend to follow users whose node degrees and HITS values (hubness or authority values) are similar to each other. The features are called as “Synchronicity” and ”Normality,” and be used in distinguishing anomalies from normal users. NFS (Network Footprint Score) [65] is an anomaly detection approach that identifies social campaigners by quantifying the likelihood of spam campaign targets. TSP-Filtering is a spammer detection approach based on isomorphic triad distribution, and I note this scheme as ‘Case 1’ experiment. Also, HH-Filtering uses feature set based on hierarchical homophily, and I note this

scheme as ‘Case 2’ experiment. For scoring algorithms (SybilRank and NFS), I set a threshold that makes the best true positive and false positive by the grid search method. Baseline methods are provided as open-source code by authors. This comparison is performed on the default parameter values of codes.

Table 5.1: Performance comparison between the proposed approaches and baseline methods.

	Twitter		**Weibo**	
	True positive (Spammer)	**False positive (Normal user)**	**True positive (Customer)**	**False positive (Normal user)**
SybilRank	33.5%	76.9%	86.0%	34.9%
NFS	44.3%	29.9%	80.0%	29.8%
CatchSync	91.5%	10.9%	75.6%	27.2%
TSP-Filtering (Case 1)	92.1%	7.9%	99.0%	0%
HH-Filtering (Case 2)	97.6%	2.4%	99.1%	0.2%

Table 5.1 is the experimental results that compare my proposed approach and baseline methods. I can observe that my proposed method outperforms all other

methods in both datasets. Most baselines performed well on Weibo dataset because spamdexing behavior occurred explicitly. However, since Twitter datasets have spammers with a variety of attack strategies, an in-depth analysis of their behavior needs to be done. SybilRank, NFS, and CatchSync use a global graph to detect spammers but show less efficiency than Case 1. Both Case 1 and Case 2 are based on ego-network, and they perform much better than three baselines. The results may indicate that the proposed features of Case 1 and Case 2 are robust such that it can effectively detect both types of spam linking attack.

Then, I compared Case 1 and Case 2 to the best baseline in terms of precision, F1 score and AUC. The result is as follows (Table 5.2).

Table 5.2: Precision, F1 score and AUC comparison in Twitter experiment.

	Precision	**F1 score**	**AUC**
CatchSync	0.903	0.904	0.951
TSP-Filtering (Case 1)	0.919	0.917	0.970
HH-Filtering (Case 2)	0.985	0.984	0.997

From Table 4.3.5, both Case 1 and Case 2 feature sets perform better than the best baseline, CatchSync. As a result, I can say that proposed features based on

structural analysis and relational semantic analysis show feasibility in spammer detection task, in terms of every evaluation metric that I measured. Finally, I evaluated the hybrid approach, Case 3. I discovered an integrated feature set outperforms the experimental results of Case 1 and Case 2. Especially, Twitter experiment shows a large improvement in the hybrid feature set. Following Table 5.3 compares performance results Case 3 to Case 1 and Case 2.

Table 5.3: Performance comparison of three feature sets.

	TSP-Filtering (Case 1)	**HH-Filtering (Case 2)**	**Hybrid Approach (Case 3)**
True positive (Spammer)	92.1%	97.6%	99.4%
False positive (Normal user)	7.9%	2.4%	0.01%

So, my proposed approach (Case 3) which utilizes both structural and relational semantic features shows true positive of 99.4% and false positive of 0.01%. This means that ego-network analysis could be the cost-effective and high-performance method for detecting online social attackers.

Additionally, I performed an evaluation on graph classification tasks with Graph Neural Network. With the state-of-the-art model [72], GNN can solve

the ego-network classification task with 94% of test accuracy in Twitter dataset. This result is similar to the evaluation result based on basic Social network-based features (In/Out degree, indegree ratio, Clustering coefficient, PageRank, Hub, Authority, and Topological sort). The result is as following table 5.4.

Table 5.4: Performance when classification with basic Social network-based features

True positive	**False positive**	**Precision**	**F1 score**
94.0%	4.5%	94.0%	94.7%

My evaluations are based on M1 status measurement. I also provide other experimental results based on M3 status measurement. Following table 5.5 is the F3 feature-based comparison. The reason why I focus on M3 is that this measurement has a spike on hierarchical gap distribution in Figure 4.2.2.2. From Table 5.5, M3 shows better performance on false positive and precision.

Table 5.5: F3 feature based comparison between M1 and M3 status measurement.

Status	**True positive**	**False positive**	**Precision**	**F1 score**
M1	90.1%	7.2%	82.6%	91.3%
M3	86.2%	4.4%	95.1%	90.5%

Chapter 6

Conclusion

Attacks on online social networks not only undermine the credibility of using SNSs, but also cause considerable economic loss to society. Circulation of abusive messages eventually hurts the credibility of the whole OSNs as an information sharing platform. Even more seriously, fraudulent connections may also be used as a vehicle for propagating social engineering schemes such as phishing that may inflict momentary damages to victims. Eventually, spammers can damage the business of OSNs because unhappy users may reduce the use of OSNs or even may quit the system. Therefore, detecting and eviction of spammers who create random connections are important in maintaining the healthy online social ecosystems. In this paper, I try to increase the design space of spam detection adding social network features. Particularly, I propose a novel spammer detection scheme that utilizes the structural analysis and unique social network characteristic called homophily. From these analyses, I found that network formation and property of online social attackers are far from normal users. As far as I know, this is the first approach that utilizes the ego-network triad distribution and homophily property in spam detection. I

conducted a performance analysis with two real-world datasets obtained from Twitter and Sina Weibo. My experimental results show that the proposed scheme improves the performance significantly. The proposed method is robust such that it can be applied for general online social network spam.

Bibliography

[1] S. M. Aghdam and N. J. Navimipour. Opinion leaders selection in the social networks based on trust relationships propagation. Karbala International Journal of Modern Science, 2(2):88–97, 2016.

[2] L. Akoglu, M. McGlohon, and C. Faloutsos. Oddball: Spotting anomalies in weighted graphs. In Advances in Knowledge Discovery and Data Mining, pages 410–421. Springer, 2010.

[3] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. The connectivity sonar: detecting site functionality by structural patterns. In Proceedings of the fourteenth ACM conference on Hypertext and hypermedia, pages 38–47. ACM, 2003.

[4] C. Avin, B. Keller, Z. Lotker, C. Mathieu, D. Peleg, and Y.-A. Pignolet. Homophily and the glass ceiling effect in social networks. In Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science, pages 41–50. ACM, 2015.

[5] P. Barber´a. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. Political Analysis, 23(1):76–91, 2015.

[6] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Web spam detection: Link-based and content-based techniques. In The

European integrated project dynamically evolving, large scale information systems (DELIS): Proceedings of the final workshop, volume 222, pages 99–113, 2008.

[7] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. A. Baeza-Yates. Link-based characterization and detection of web spam. In AIRWeb, pages 1–8, 2006.

[8] A. A. Benczu´r, K. Csalog´any, and T. Sarl´os. Link-based similarity search to fight web spam. In In AIRWEB. Citeseer, 2006.

[9] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), volume 6, page 12, 2010.

[10] C. A. Bliss, I. M. Kloumann, K. D. Harris, C. M. Danforth, and P. S. Dodds. Twitter reciprocal reply networks exhibit assortativity with respect to happiness. Journal of Computational Science, 3(5):388–397, 2012.

[11] J. Bollen, B. Gon¸calves, G. Ruan, and H. Mao. Happiness is assortative in online social networks. Artificial life, 17(3):237–251, 2011.

[12] Y. Boshmaf, D. Logothetis, G. Siganos, J. Ler´ıa, J. Lorenzo, M. Ripeanu, and K. Beznosov. Integro: Leveraging victim prediction for robust fake account detection in osns. In NDSS, volume 15, pages 8–11, 2015.

[13] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro. Aiding the detection of fake accounts in large scale social online services. In Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, pages 15–15. USENIX Association, 2012.

[14] Q. Cao, X. Yang, J. Yu, and C. Palow. Uncovering large groups of active malicious accounts in online social networks. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pages 477–488. ACM, 2014.

[15] D. Cartwright and F. Harary. Structural balance: a generalization of heider's theory. Psychological review, 63(5):277, 1956.

[16] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the web topology. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 423–430. ACM, 2007.

[17] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. ICWSM, 10(10-17):30, 2010.

[18] G. Danezis and P. Mittal. Sybilinfer: Detecting sybil nodes using social networks. In NDSS. San Diego, CA, 2009.

[19] J. R. Douceur, A. Adya, W. J. Bolosky, D. Simon, and M. Theimer. Reclaiming space from duplicate files in a serverless distributed file system. In Distributed Computing Systems, 2002. Proceedings. 22nd International Conference on, pages 617–624. IEEE, 2002.

[20] I. Drost and T. Scheffer. Thwarting the nigritude ultramarine: Learning to identify link spam. In European Conference on Machine Learning, pages 96–107. Springer, 2005.

[21] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna. Compa: Detecting compromised accounts on social networks. In NDSS, 2013.

[22] S. Fakhraei, J. Foulds, M. Shashanka, and L. Getoor. Collective spammer detection in evolving multi-relational social networks. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1769–1778. ACM, 2015.

[23] Y. Fan, Y. Zhang, Y. Ye, and X. Li. Automatic opioid user detection from twitter: Transductive ensemble built on different meta-graph based similarities over heterogeneous information network. In IJCAI, pages 3357– 3363, 2018.

[24] M. Fazeen, R. Dantu, and P. Guturu. Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and contextindependent approaches. Social Network Analysis and Mining, 1(3):241– 254, 2011.

[25] L. C. Freeman. Centered graphs and the structure of ego networks. Mathematical Social Sciences, 3(3):291–304, 1982.

[26] H. Gao, Y. Yang, K. Bu, Y. Chen, D. Downey, K. Lee, and A. Choudhary. Spam ain't as diverse as it seems: throttling osn spam with templates

underneath. In Proceedings of the 30th Annual Computer Security Applications Conference, pages 76–85. ACM, 2014.

[27] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi. Understanding and combating link farming in the twitter social network. In Proceedings of the 21st international conference on World Wide Web, pages 61–70. ACM, 2012.

[28] N. Z. Gong, M. Frank, and P. Mittal. Sybilbelief: A semi-supervised learning approach for structure-based sybil detection. Information Forensics and Security, IEEE Transactions on, 9(6):976–987, 2014.

[29] P. Graham. A plan for spam, 2002.

[30] S. Gupta, A. Khattar, A. Gogia, P. Kumaraguru, and T. Chakraborty. Collective classification of spam campaigners on twitter: A hierarchical meta-path based approach. In Proceedings of the 2018 World Wide Web Conference on World Wide Web, pages 529–538. International World Wide Web Conferences Steering Committee, 2018.

[31] Z. Gy¨ongyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, pages 576–587. VLDB Endowment, 2004.

[32] F. Heider. Social perception and phenomenal causality. Psychological review, 51(6):358, 1944.

[33] F. Heider. Attitudes and cognitive organization. The Journal of psychology, 21(1):107–112, 1946.

[34] I. Himelboim, S. McCreery, and M. Smith. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. Journal of computer-mediated communication, 18(2):154– 174, 2013.

[35] J. Hovold. Naive bayes spam filtering using word-position-based attributes. In CEAS, pages 41–48, 2005.

[36] S. Jeong, J. Lee, J. Park, and C.-k. Kim. The social relation key: A new paradigm for security. Information Systems, 71:68–77, 2017.

[37] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang. Catchsync: catching synchronized behavior in large directed graphs. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 941–950. ACM, 2014.

[38] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi. The social world of content abusers in community question answering. In Proceedings of the 24th International Conference on World Wide Web, pages 570–580. International World Wide Web Conferences Steering Committee, 2015.

[39] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 46(5):604–632, 1999.

[40] J. M. Kleinberg. Hubs, authorities, and communities. ACM computing surveys (CSUR), 31(4es):5, 1999.

[41] V. Krishnan and R. Raj. Web spam detection with anti-trust rank. In AIRWeb, volume 6, pages 37–40, 2006.

[42] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In Proceedings of the 19th international conference on World wide web, pages 591–600. ACM, 2010.

[43] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1361–1370. ACM, 2010.

[44] X. Li, M. Zhang, Y. Liu, S. Ma, Y. Jin, and L. Ru. Search engine click spam detection based on bipartite graph propagation. In Proceedings of the 7th ACM international conference on Web search and data mining, pages 93–102. ACM, 2014.

[45] J. Martinez-Romo and L. Araujo. Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Systems with Applications, 40(8):2992–3000, 2013.

[46] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. Annual review of sociology, 27(1):415–444, 2001.

[47] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. Science, 303(5663):1538–1542, 2004.

[48] D. O'Callaghan, M. Harrigan, J. Carthy, and P. Cunningham. Network analysis of recurring youtube spam campaigns. In Sixth International AAAI Conference on Weblogs and Social Media, 2012.

[49] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[50] D. Sculley and G. M. Wachman. Relaxed online svms for spam filtering. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 415–422. ACM, 2007.

[51] L. Shi, S. Yu, W. Lou, and Y. T. Hou. Sybilshield: An agent-aided social network-based sybil defense among multiple communities. In INFOCOM, 2013 Proceedings IEEE, pages 1034–1042. IEEE, 2013.

[52] T. Tian, J. Zhu, F. Xia, X. Zhuang, and T. Zhang. Crowd fraud detection in internet advertising. In Proceedings of the 24th International Conference on World Wide Web, pages 1100–1110. International World Wide Web Conferences Steering Committee, 2015.

[53] N. Tran, J. Li, L. Subramanian, and S. S. Chow. Optimal sybil-resilient node admission control. In INFOCOM, 2011 Proceedings IEEE, pages 3218–3226. IEEE, 2011.

[54] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In Proceedings of the 23rd USENIX Security Symposium (USENIX Security), 2014.

[55] A. H. Wang. Don't follow me: Spam detection in twitter. In Security and Cryptography (SECRYPT), Proceedings of the 2010 International Confer ence on, pages 1–10. IEEE, 2010.

[56] S. Wasserman and K. Faust. Social network analysis: Methods and applications, volume 8. Cambridge university press, 1994.

[57] M. Weber. Class, status, party. 1993.

[58] W. Wei, F. Xu, C. C. Tan, and Q. Li. Sybildefender: Defend against sybil attacks in large social networks. In INFOCOM, 2012 Proceedings IEEE, pages 1951–1959. IEEE, 2012.

[59] B. Wu and B. D. Davison. Identifying link farm spam pages. In Special interest tracks and posters of the 14th international conference on World Wide Web, pages 820–829. ACM, 2005.

[60] B. Wu, V. Goel, and B. D. Davison. Topical trustrank: Using topicality to combat web spam. In Proceedings of the 15th international conference on World Wide Web, pages 63–72. ACM, 2006.

[61] K. Wu, S. Yang, and K. Q. Zhu. False rumors detection on sina weibo by propagation structures. In 2015 IEEE 31st international conference on data engineering, pages 651–662. IEEE, 2015.

[62] L. Wu, X. Hu, F. Morstatter, and H. Liu. Adaptive spammer detection with sparse group modeling. In Eleventh International AAAI Conference on Web and Social Media, 2017.

[63] W. Xie, C. Li, F. Zhu, E.-P. Lim, and X. Gong. When a friend in twitter is a friend in life. In Proceedings of the 4th Annual ACM Web Science Conference, pages 344–347. ACM, 2012.

[64] S. Yardi, D. Romero, G. Schoenebeck, et al. Detecting spam in a twitter network. First Monday, 15(1), 2009.

[65] J. Ye and L. Akoglu. Discovering opinion spammer groups by network footprints. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 267–282. Springer, 2015.

[66] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao. Sybillimit: A nearoptimal social network defense against sybil attacks. In Security and Privacy, 2008. SP 2008. IEEE Symposium on, pages 3–17. IEEE, 2008.

[67] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. Sybilguard: defending against sybil attacks via social networks. ACM SIGCOMM Computer Communication Review, 36(4):267–278, 2006.

[68] D. Yuan, G. Li, Q. Li, and Y. Zheng. Sybil defense in crowdsourcing platforms. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 1529–1538. ACM, 2017.

[69] Y. Zhang and J. Lu. Discover millions of fake followers in weibo. Social Network Analysis and Mining, 6(1):16, 2016.

[70] H. Zheng, M. Xue, H. Lu, S. Hao, H. Zhu, X. Liang, and K. Ross. Smoke screener or straight shooter: Detecting elite sybil attacks in user-review social networks. arXiv preprint arXiv:1709.06916, 2017.

[71] B. Zhou, Y. Yao, and J. Luo. Cost-sensitive three-way email spam filtering. Journal of Intelligent Information Systems, 42(1):19–45, 2014.

[72] Chen, Ting, Song Bian, and Yizhou Sun. "Are Powerful Graph Neural Nets Necessary? A Dissection on Graph Classification." arXiv preprint arXiv:1905.04579 (2019).

국 문 초 록

최근 우리는 Facebook, Twitter, Weibo, LinkedIn 등의 다양한 사회 관계망 서비스가 폭발적으로 성장하는 현상을 목격하였다. 하지만 사회 관계망 서비스가 개인과 개인간의 관계 및 커뮤니티 형성과 뉴스 전파 등의 여러 이점을 제공해 주고 있는데 반해 반갑지 않은 현상 역시 발생하고 있다. 스패머들은 사회 관계망 서비스를 동력 삼아 스팸을 매우 빠르고 넓게 전파하는 식으로 악용하고 있다. 스팸은 수신자가 원치 않는 메시지들을 일컫는데 이는 서비스의 신뢰도와 안정성을 크게 손상시킨다. 따라서, 스패머를 탐지하는 것이 현재 소셜 미디어에서 매우 긴급하고 중요한 문제가 되었다. 이 논문은 대표적인 사회 관계망 서비스들 중 Twitter 와 Weibo 에서 발생하는 스패밍을 다루고 있다. 이러한 유형의 스패밍들은 불특정 다수에게 메시지를 전파하는 대신에, 많은 일반 사용자들을 '팔로우(구독)'하고 이들로부터 '맞 팔로잉(맞 구독)'을 이끌어 내는 것을 목적으로 하기도 한다. 때로는 link farm 을 이용해 특정 계정의 팔로워 수를 높이고 명시적 영향력을 증가시키기도 한다. 스패머의 온라인 관계망이 일반 사용자의 온라인 사회망과 다를 것이라는 가정 하에, 나는 스패머들을 포함한 일반적인 온라인 사회망 공격자들을 탐지하는 분류 방법을 제시한다. 나는 먼저 개인 사회망 내 사회 관계에 주목하고 두 가지 종류의 분류 특성을 제안하였다. 이들은 개인 사회망의 Triad Significance Profile (TSP)에 기반한 구조적 특성과 Hierarchical homophily 에 기반한 관계 의미적 특성이다. 실제 Twitter 와 Weibo 데이터셋에 대한 실험 결과는 제안한 방법이 매우

실용적이라는 것을 보여준다. 제안한 특성들은 전체 네트워크를 분석하지 않아도 개인 사회망만 분석하면 되기 때문에 scalable 하게 측정될 수 있다. 나의 성능 분석 결과는 제안한 기법이 기존 방법에 비해 true positive 와 false positive 측면에서 우수하다는 것을 보여준다.

주요어 : 온라인 소셜 네트워크, 스팸, 개인 사회망, 동류성, 지위, 이상 탐지, 사용자 특성 분석

학 번 : 2014-21789