Robust Spammer Detection in Microblogs: Leveraging User Carefulness

HAO FU, University of Science and Technology of China
XING XIE and YONG RUI, Microsoft Research
NEIL ZHENQIANG GONG, Iowa State University
GUANGZHONG SUN and ENHONG CHEN, University of Science and Technology of China

Microblogging Web sites, such as Twitter and Sina Weibo, have become popular platforms for socializing and sharing information in recent years. Spammers have also discovered this new opportunity to unfairly overpower normal users with unsolicited content, namely social spams. Although it is intuitive for everyone to follow legitimate users, recent studies show that both legitimate users and spammers follow spammers for different reasons. Evidence of users seeking spammers on purpose is also observed. We regard this behavior as useful information for spammer detection. In this article, we approach the problem of spammer detection by leveraging the "carefulness" of users, which indicates how careful a user is when she is about to follow a potential spammer. We propose a framework to measure the carefulness and develop a supervised learning algorithm to estimate it based on known spammers and legitimate users. We illustrate how the robustness of the detection algorithms can be improved with aid of the proposed measure. Evaluation on two real datasets from Sina Weibo and Twitter with millions of users are performed, as well as an online test on Sina Weibo. The results show that our approach indeed captures the carefulness, and it is effective for detecting spammers. In addition, we find that our measure is also beneficial for other applications, such as link prediction.

Categories and Subject Descriptors: H.2.7 [**Database Management**]: Database Administration—*Security, Integrity, and Protection*; H.2.8 [**Database Management**]: Database Applications—*Data Mining*

General Terms: Algorithms, Experimentation, Security

Additional Key Words and Phrases: Spammer detection, social network, microblog

ACM Reference Format:
Hao Fu, Xing Xie, Yong Rui, Neil Zhenqiang Gong, Guangzhong Sun, and Enhong Chen. 2017. Robust spammer detection in microblogs: Leveraging user carefulness. ACM Trans. Intell. Syst. Technol. 8, 6, Article 83 (August 2017), 31 pages.
DOI: http://dx.doi.org/10.1145/3086637

1. INTRODUCTION

In recent years, microblogging Web sites, such as Twitter and Sina Weibo, have gained increasing popularity. With rapidly growing influence among users, they have become a universal platform for sharing personal experience, marketing, mass media, and public relationship. Similarly to other online social networking Web sites [Heymann et al. 2007], spammers have discovered microblogging as an appealing platform to spread

This work was supported in part by the Youth Innovation Promotion Association of the Chinese Academy of Sciences and the National Science Foundation for Distinguished Young Scholars of China (grant 61325010).

Authors' addresses: H. Fu (corresponding author), G. Sun, and E. Chen, University of Science and Technology of China, No. 96, JinZhai Road Baohe District, Hefei, Anhui, 230026, PR China; emails: fuch@mail.ustc.edu.cn, {gzsun, cheneh}@ustc.edu.cn; X. Xie and Y. Rui, Microsoft Research, No. 5, Danling Street, Haidian District, Beijing 100080, PR China; emails: {xingx, yongrui}@microsoft.com; N. Z. Gong, Iowa State University, 104 Marston Hall, 533 Morrill Road, Ames, IA 50011-2151, USA; email: neilgong@iastate.edu.

DOI: http://dx.doi.org/10.1145/3086637

spams with fake accounts. Spams not only annoy users but also lead to financial loss and privacy risks of users. Robust detection of spammers, which improves the quality of user experience and social systems, is certainly necessary.

One of the main challenges is that spammers are upgrading their spamming strategies rapidly to race with the development of detection systems. A detection system that is able to capture most of the spammers one month may fail the next month. For example, it has long been a common practice for email server administrators to update spam filters frequently. To camouflage themselves, spammers may manipulate profiles, tweets, and social relationships of their accounts. Tweets and profiles have been shown to be good information sources for detection [Stringhini et al. 2010; Hu et al. 2013], but they can be faked by spammers if they wish. In addition, access to the content is sometimes restricted due to privacy concerns [Zhu et al. 2012].

In a microblogging Web site, a user decides who to follow based on her own knowledge. Although spammers can simulate normal link patterns between their fake accounts, they can hardly affect the decisions of legitimate users. We regard such links as a robust information source for spammer detection. In this article, we focus on detecting spammers based on links.

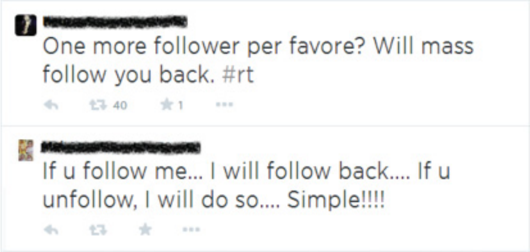
It is intuitive and necessary for spammers to follow legitimate users so that they can attract attention from legitimate users and spread spams. However, conflicting observations have been made on whether spammers would connect to other spammers. Zhu et al. [2012] found that spammers are separated on Renren, which is a Facebook-like social network. Yang et al. [2012] had an opposite finding on Twitter, where spammers tend to be interconnected, possibly trying to simulate normal link patterns. Consequently, different algorithms for spammer detection were proposed for the two networks [Zhu et al. 2012; Hu et al. 2013].

It is commonly agreed that legitimate users favor only other legitimate users and do not follow others at random. For example, Weng et al. [2010] found that the presence of reciprocity on Twitter can be explained by the theory of “homophily.” Users sharing similar topics are more likely to follow each other reciprocally. Hopcroft et al. [2011] showed strong evidence of the structural balance among reciprocal relationships—that is, users with common friends of reciprocal ties have a tendency to follow each other. The preceding findings indicate that some users do follow others “seriously.” However, evidence of legitimate users following spammers was also found. Ghosh et al. [2012] discovered that a small fraction of users, namely social capitalists, are seeking to increase their social capital by following back anyone who follows them. Yang et al. [2012] also observed similar users, who in turn aid spammers to spread spams and avoid detection.

The preceding discussion implies that the intention of a “follow” action (favoring legitimate users or spammers) varies among users. A well-intentioned user is expected to follow legitimate users seriously, but she may also follow spammers inadvertently. A malicious user is expected to cooperate with spammers, but she may also need to follow some legitimate users to appear normal. This leads to an interesting question: can we measure how serious a user is when she is trying to follow someone?

In the context of spammer detection, we refer to this property as the *carefulness*, which indicates how careful a user is when she is trying to avoid spammers. The carefulness is able to characterize the following behaviors of users. A *careful* user typically follows only legitimate users and always manages to avoid spammers. A *careless* user could be either well intentioned or malicious, but she shows no particular preference toward legitimate users or spammers. An extremely *malicious* user typically follows only spammers but pays no attention to legitimate users.

It should be noted that many previous works on spammer detection [Chirita et al. 2005; Cao et al. 2012; Xue et al. 2013] assume that legitimate users favor only other



[redacted]
One more follower per favore? Will mass follow you back. #rt

[redacted]
If u follow me… I will follow back…. If u unfollow, I will do so…. Simple!!!!

Fig. 1. Example of casual users when following others.

legitimate users. We avoid such assumptions by leveraging the proposed carefulness. For example, as shown in Figure 1, the users themselves are legitimate but careless. They follow back anyone who follows them, so they are potentially following spammers.

Given the carefulness of users, we are interested in how it can be leveraged to aid spammer detection. Previous works have observed that the behavior of spammers varies among different networks. Spammers in one network may form tightly connected communities [Yang et al. 2012] but spread in the wild in another network [Yang et al. 2011; Zhu et al. 2012]. We also have a similar observation in our dataset, which contains two social graphs from Twitter and Sina Weibo, respectively. Consequently, different detection algorithms are proposed to capture the distinct behavior of spammers. A traditional approach is extracting graph-based features and training a classifier [Benevenuto et al. 2010]. A limitation of this approach is that the features are mainly based on one's ego network (e.g., degrees and clustering coefficients). When spammers form fake communities intentionally, these features can be easily manipulated, making them less distinguishable. To address this issue, the propagation approach is developed to leverage the community structure of spammers [Yu et al. 2006]. It starts with a few known spammers and propagates scores via links, aiming to discover communities of spammers. It assumes that spammers are tightly connected and have fewer links with legitimate users. Unfortunately, this assumption is true only in some networks [Yang et al. 2011].

The preceding discussion shows that existing detection algorithms are not robust regarding the difference in spammers' behavior. An algorithm that works well in one network may not work in another network. We investigate how the proposed carefulness is incorporated with existing detection algorithms, aiming to improve the robustness of the detection algorithms.

In this article, we try to answer the following two questions:

—Can we measure how serious a user is when she is trying to follow someone?
—How can the carefulness be leveraged to aid spammer detection?

We make the following contributions to answer the two questions:

—We propose a framework to quantify the carefulness of users and develop a supervised learning algorithm to estimate it based on known spammers and legitimate users.
—We review previously proposed algorithms for spammer detection and illustrate how the carefulness is incorporated to improve the detection.
—We evaluate our method on two real datasets from Sina Weibo and Twitter consisting of millions of users. Our results show that our method is able to characterize user behavior in terms of the carefulness. With the help of the proposed carefulness, existing algorithms can be enhanced to detect spammers robustly.

—We characterize spammers and their network neighbors on Twitter and Sina Weibo. Significant difference is observed, which explains why an algorithm may not work equally well in the two networks.
—We illustrate how other applications (e.g., link prediction) can benefit from the proposed carefulness.

In the rest of this article, we first review related works and discuss the difference (Section 2). After giving a concrete formulation of our problem (Section 3), we start by introducing the definition and the learning algorithm of the carefulness (Section 4). We then discuss how to incorporate the carefulness to improve spammer detection (Section 5). A description and observation of our dataset is presented in Section 6. Evaluation of our approach is presented in Section 7. Several technical issues and other applications are discussed in Section 8. Finally, we conclude our results and discuss future works based on the proposed method (Section 9).

2. RELATED WORK

Spammer detection in social networks, such as email systems [Boykin and Roychowdhury 2005; Chirita et al. 2005] and SMS networks [Xu et al. 2012], has been widely studied for many years. In recent years, spammers in microblogging Web sites have attracted increasing attention from researchers and developers. Many works focus on characterizing abnormal or spamming behaviors in various aspects [Yardi et al. 2009; Grier et al. 2010; Thomas et al. 2011; Ghosh et al. 2012; Yang et al. 2012]. Another major topic is detecting spammers based on tweets, network structure, or both. In this article, we focus on detecting spammers based on network structure.

To the best of our knowledge, three approaches have been developed for detection: feature based, propagation over the network, and matrix factorization.

2.1. Feature Based

A traditional approach is extracting various features from the network and training a classifier for detection [Benevenuto et al. 2010; Yang et al. 2011]. For instance, Benevenuto et al. [2010] studied the problem of spammer detection on Twitter. They analyzed the tweet content and user social activities on Twitter, from which several features were extracted for detection. As the features mostly capture the local structure of nodes (e.g., degrees only count the number of one-hop neighbors), they can be easily manipulated by spammers. Spammers can simulate normal link patterns between fake accounts, making them indistinguishable from normal accounts. We find that the proposed carefulness can mitigate such manipulation.

2.2. Propagation

Propagation-based methods assume some kind of correlation between a pair of follower and followee (e.g., being similar or dissimilar). Scores of being spamming or legitimate are propagated via the links of graph, according to certain probabilistic models (e.g., random walk and probabilistic graphical models).

Several works adapt random walk models to rank spammers based on network structure. The general idea is that legitimate users create links only to other legitimate users. Users who receive more links from legitimate users are more likely to be legitimate. Gyöngyi et al. [2004] proposed TrustRank to detect Web spams. TrustRank is initiated with a set of known good Web sites as seeds and then propagates the scores with biases. Chirita et al. [2005] proposed ranking the reputation of email senders with a variant of PageRank. Xue et al. [2013] utilized friend requests to enhance the detection. Many works are also based on similar models, such as SybilGuard [Yu et al. 2006], SybilLimit [Yu et al. 2008], SybilInfer [Danezis and Mittal 2009], and SybilRank [Cao et al. 2012]. Boshmaf et al. [2015] proposed detecting spammers by predicting victims who follow spammers accidentally first. Their approach requires manually labeling a training dataset of victims and nonvictims, which is hard to obtain in practice. Moreover, predicting victims relies on user profiles and content; however, our approach only requires network structure. Some other algorithms leverage probabilistic graphical models, such as SybilBelief [Gong et al. 2014a], SybilFrame [Gao et al. 2015], and SybilSCAR [Wang et al. 2017]. The correlation between connected users are modeled in a probabilistic manner, which provides certain flexibility and yields better performance.

Many of these algorithms assume that neighboring nodes are similar in terms of being spamming or legitimate. However, as shown in Yang et al. [2011] and Zhu et al. [2012], spammers do not form communities in some social networks, so these algorithms are not applicable there. Random walk–based methods assume that legitimate users favor only other legitimate users. However, the case that legitimate users follow spammers [Ghosh et al. 2012; Yang et al. 2012], which occurs quite often, is not considered. To address this issue, we propose the carefulness to characterize such behavior separately.

A recently proposed algorithm, namely SybilBelief [Gong et al. 2014a], models the cases where neighboring nodes are similar or dissimilar. SybilBelief is a state-of-the-art propagation-based algorithm that has been shown to outperform many existing algorithms. We will illustrate how it can be enhanced with the proposed carefulness.

2.3. Matrix Factorization

Recently, matrix factorization techniques have been employed to detect spammers. The general idea is modeling social network users with a set of latent factors. Online activities and network structure are a consequence of the interactions between the latent factors. Spammers and legitimate users are different in terms of the latent factors. Hu et al. [2013, 2014] proposed a family of matrix factorization methods for this problem. They assumed that neighboring users tend to be both spammers or legitimate users and made use of the content of tweets. Zhu et al. [2012] also employed a matrix factorization approach for Facebook-like social networks. They made a different assumption about neighboring users: whereas legitimate users are interconnected, spammers are apart from each other. Their approach does not rely on the content of posts or profiles but requires the records of user activities.

These methods require additional information other than the network structure and seem to underestimate the knowledge of legitimate users. As shown in this article, certain hidden traits of users (e.g., the carefulness) are very useful for the detection. Additionally, we do not make any assumption on whether spammers are connected with or apart from each other. As the matrix factorization methods require additional information, we only discuss how the first two approaches can leverage the proposed carefulness in this work.

3. PRELIMINARY

Before introducing our approach in detail, we give a concrete definition of notions and the problem.

3.1. Definition

We model users and their social ties in a microblogging Web site as a directed social graph $G = (V, E)$. Every node in V corresponds to a unique user on the Web site. A directed link $(u, v) \in E$ is presented in the graph if and only if the user u is following v. The links (u, v) and (v, u) may both exist if the users are following each other reciprocally. We denote the followers of a user v as a set $N_I(v) = \{u | (u, v) \in E\}$. The followees of

Table I. Table of Notations

Symbol	Section	Meaning
$G = (V, E)$	Section 3.1	Social graph with node set V and edge set E
$N_I(u)$	Section 3.1	Set of user u's followers
$N_O(u)$	Section 3.1	Set of user u's followees
$N_R(u)$	Section 3.1	Set of user u's friends
$f(u)$	Section 4.2	Carefulness of user u
Y_u	Section 4.2	Label of user u being spamming or legitimate
X_u	Section 4.3	Feature set of user u
D	Section 4.3	Training set
$\mathbf{w}$	Section 4.3	Parameter of the carefulness model
$g(u)$	Section 4.3	Prediction on if user u is a spammer based on the carefulness of followers
$L(\mathbf{w})$	Section 4.3	Loss function of the carefulness model
$C_O(u)$, $C_R(u)$	Section 5.1.2	Clustering coefficients
$PR(u)$	Section 5.1.3	PageRank
$RPR(u)$	Section 5.1.3	Reversed PageRank
$\phi_v(Y_v)$, $\varphi_{uv}(Y_u, Y_v)$	Section 5.2	Potential functions of SybilBelief
α, β	Section 5.2	Parameters of adjusted edge potential

a user u are represented as a set $N_O(u) = \{v | (u, v) \in E\}$. Additionally, friends of the user v (those who have reciprocal relations with v) are denoted as $N_R(v) = N_I(v) \cap N_O(v)$. In the rest of this article, we would use the terms *user* and *node* interchangeably.

Important notations are listed in Table I. Detailed definitions will be presented when the notations are first used.

3.2. Problem Formulation

Given a social graph G, our first goal is learning a function $f(u)$ that estimates the carefulness of the user u when she is about to follow someone else. A high value of $f(u)$ indicates that u favors legitimate users and avoid spammers carefully. A low value implies that u follows spammers, which appears somewhat careless or even malicious. Our second goal is detecting spammers based on the graph structure and the learned carefulness $f(u)$.

4. MINING CAREFULNESS

In this section, we first discuss how often a user would follow spammers. We then propose a framework to model the carefulness $f(u)$. Based on the proposed model, we introduce an algorithm that learns $f(u)$ from known spammers and legitimate users.

4.1. Spamming Followees

Thus far, we know that it is possible for both legitimate users and spammers to follow spammers, but how often does it happen? We used Sina Weibo,[1] which is one of the most popular microblogging Web sites in China, to seek answers to this question.

Our dataset contains 3.5 million users, and 2,000 users are manually identified as legitimate user or spammer (see Section 7.2). We consider the fraction of spamming followees as a case study here. Due to the limited number of known spammers, we consider the fraction of suspended users instead, which can be massively crawled from the Web site. In our dataset, 8.4% of users are suspended by Sina Weibo mainly due to abusive activities. If a user follows others at random, the expected fraction of suspended followees would be 8.4%. Among the identified 2,000 users, 71.8% of legitimate users and 68.9% of spammers follow at least one suspended user. More importantly, 11.5%

[1] http://www.weibo.com.

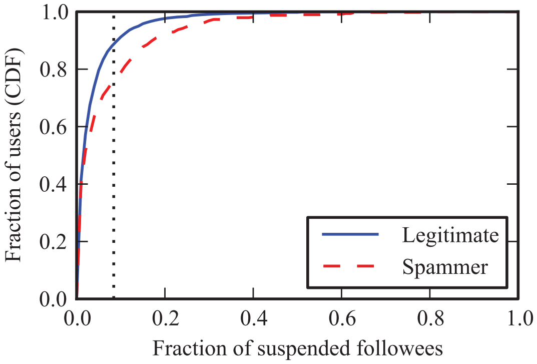


Fig. 2. Cumulative distributions of the fraction of suspended followees for legitimate users and spammers. The vertical line denotes the expected fraction if a user follows others at random.

of legitimate users and 23.7% of spammers follow more suspended users than random (Figure 2). Note that the fraction of spamming followees is underestimated here because more spammers are not suspended yet.

The observation shows that legitimate users follow spammers quite often. We find that most legitimate users who follow more spammers than random are marketers. A possible explanation is that they are cooperating with spammers to promote their products. We notice that hijacked accounts may also follow more spammers. This is observed via tweets posted by the real users complaining about the hijacking after they reclaimed their accounts. Spammers follow significantly more spammers than legitimate users do, which implies that spammers are trying to camouflage themselves by increasing the number of followers on purpose. In addition, the majority of followees are still legitimate for both legitimate users and spammers. This is expected because most users are legitimate. This observation completes our discussion about the behavior of following spammers.

4.2. Carefulness

We define the carefulness of the user u as the probability of identifying another user as a legitimate user or a spammer correctly. To simplify the problem, we assume that the probability only depends on the user u, so it is denoted as a function $f(u)$.

The carefulness is not directly accessible, so we have to estimate it via other observable information. The preceding observation suggests that it can be inferred from one's followees. With a handful of spammers identified by experts, we can build the connection between the labels and the carefulness. We use the variable Y_v to denote the label of v. We define $Y_v = 1$ if v is a spammer or $Y_v = 0$ otherwise.

When a user v comes, the user u may decide whether to follow v based on her knowledge of v. User u is assumed to follow only users that she considers legitimate. However, if v is considered legitimate, u does not necessarily have to follow v. For example, it is also determined by various properties of the network, such as proximity [Liben-Nowell and Kleinberg 2003], homophily [Weng et al. 2010], and structural balance [Hopcroft et al. 2011]. Given that u considers v legitimate, we define $r(u, v)$ as the conditional probability of actually forming a directed link from u to v. Given that v is a legitimate user or a spammer, we have the probability of a "follow" action as

$$\begin{aligned} P((u, v) \in E | Y_v = 0) &= f(u) r(u, v), \\ P((u, v) \in E | Y_v = 1) &= (1 - f(u)) r(u, v). \end{aligned} \tag{1}$$

It can be shown that the proposed model is able to capture the following typical types of users:

—*Careful* users who always follow legitimate users and never make mistakes ($f(u) = 1$)
—*Careless* users who do not make effort to identify spammers, showing no particular preference toward legitimate users or spammers ($f(u) = 1/2$)
—*Malicious* users who always seek spammers ($f(u) = 0$).

Finding a proper estimation of $r(u, v)$ is complicated. We try to avoid it by focusing on only existing links. By applying Bayes' rule, we have

$$\begin{aligned} P(Y_v = 1 | (u, v) \in E) \\ &= \frac{P((u, v) \in E | Y_v = 1) P(Y_v = 1)}{\sum_{y \in \{0,1\}} P((u, v) \in E | Y_v = y) P(Y_v = y)} \\ &= \frac{(1 - f(u)) P(Y_v = 1)}{f(u) P(Y_v = 0) + (1 - f(u)) P(Y_v = 1)}. \end{aligned} \tag{2}$$

Given an existing link (u, v), Equation (2) shows that $f(u)$ only depends on whether the followees are legitimate, which means that we can simply ignore $r(u, v)$. Ideally, if we manage to identify a sufficient number of legitimate users and spammers among u's followees, we may easily estimate $f(u)$ according to Equation (2). This is infeasible due to the incredible amount of manual work. As a result, we need to develop an approach that requires fewer known spammers and legitimate users.

4.3. Our Approach

We employ a supervised learning approach to infer the carefulness based on only a few known spammers and legitimate users. We define $f(u)$ as a function of features $X_u = (x_{u1}, x_{u2}, \ldots, x_{uk})$ associated with u:

$$f(u) = \frac{1}{1 + \exp\left(-\sum_{i=0}^{k} w_i x_{ui}\right)}. \tag{3}$$

A dummy feature $x_{u0} = 1$ is included to make w_0 an intercept. In this article, we only focus on structural features (e.g., degrees) and leave the use of user profiles and tweets for future work. In our experiments, we use $k = 9$ features, including the number of followees/followers/reciprocal relations, response rate, follow-back rate, two versions of clustering coefficient, PageRank, and reversed PageRank. Detailed definitions of the features can be found in Section 5.1.

The logistic function $f(u) \in (0, 1)$ is widely used to estimate probabilities in machine learning algorithms (e.g., logistic regression and artificial neural networks). We find it a good choice for this problem in our initial experiments. This definition actually assumes a correlation between graph structure and the carefulness. For example, it is unlikely for a user to examine thousands of followees if she has that many, so we may consider the user somewhat careless.

We propose the function $g(v)$ as a prediction on if v is a spammer based on the carefulness of followers. The function $g(v)$ should be continuous and differentiable so that the learning process can be easily formulated as an optimization problem similarly to most machine learning algorithms. The function $g(v)$ should be negatively associated with the carefulness of v's followers. For example, if all followers of v are very careful ($f(u) = 1$), it is a strong evidence for v being legitimate. In this case, we shall define the value of $g(v)$ as 0. When some of the followers are found careless, a larger value should be assigned to $g(v)$. In an extreme case that all followers are malicious ($f(u) = 0$), we have to assume that v is a spammer. As malicious users are seeking spammers on purpose, it is unlikely for a legitimate user to gain so much attention from them.

Regarding the preceding requirements, we find that the average of $P(Y_v = 1|(u, v) \in E)$ for $u \in N_I(v)$ is a good choice for $g(v)$. We consider the prior probability $P(Y_v = 1)$ as a weak prediction on if v is a spammer. It can be estimated in multiple ways. For the sake of simplicity, we approximate $P(Y_v = 1) = p_s$ as the fraction of spammers in the training set. The function $g(v)$ is thus defined as

$$\begin{aligned} g(v) &= \frac{1}{|N_I(v)|} \sum_{u \in N_I(v)} P(Y_v = 1|(u, v) \in E) \\ &= 1 - \frac{1}{|N_I(v)|} \cdot \sum_{u \in N_I(v)} \frac{1}{1 + \frac{p_s}{1-p_s} \exp\left(-\sum_{i=0}^{k} w_i x_{ui}\right)}. \end{aligned} \tag{4}$$

Given a set D of labeled users, our goal is to determine the value of $\mathbf{w}$ such that minimizes the difference between the prediction $\hat{y}_v = g(v)$ and the actual label y_v. We quantify the difference with the squared error, and a regularization term is added to avoid overfitting:

$$\arg\min_{\mathbf{w}} L(\mathbf{w}) = \frac{1}{2} \sum_{v \in D} \left(g(v) - y_v\right)^2 + \frac{\lambda}{2} \sum_{i=0}^{k} w_i^2. \tag{5}$$

The parameter λ trades off between the complexity and the fitness of the model. In our initial experiments, we find that $\lambda = 1$ yields good detection performance in most cases, so we use this value throughout our experiments. We discuss other choices of the loss function in Section 8.1.

4.4. Training the Model

The learning process can be stated as an optimization problem that minimizes the loss function $L(\mathbf{w})$. We first have the gradient of $L(\mathbf{w})$ as

$$\frac{\partial L(\mathbf{w})}{\partial w_i} = \sum_{v \in D} \left(g(v) - y_v\right) \frac{\partial g(v)}{\partial w_i} + \lambda w_i. \tag{6}$$

Taking the derivative of $g(v)$ gives

$$\begin{aligned} \frac{\partial g(v)}{\partial w_i} &= -\frac{1}{|N_I(v)|} \sum_{u \in N_I(v)} \frac{\partial d(u)}{\partial w_i} \\ &= -\frac{1}{|N_I(v)|} \sum_{u \in N_I(v)} d(u)(1 - d(u)) \cdot x_{ui}, \end{aligned} \tag{7}$$

where $d(u)$ is defined as

$$d(u) = \frac{1}{1 + \frac{p_s}{1-p_s} \exp\left(-\sum_{i=0}^{k} w_i x_{ui}\right)}. \tag{8}$$

We apply the BFGS algorithm to solve the optimization problem. All features are standardized for better convergence. The algorithm may get stuck in a local minimum, so we repeat the algorithm several times with different starting points to find a good solution. We terminate the iteration when the relative improvement is less than 0.1%. In our experiments, the BFGS algorithm converges in fewer than 20 iterations. Finally, we calculate the carefulness $f(u)$ for all users with the learned parameter $\mathbf{w}$. We discuss how it is leveraged to detect spammers in the next section.

5. DETECTING SPAMMERS

Given the carefulness $f(u)$, it is still unclear how it can be leveraged to detect spammers on a microblogging Web site. A trivial way is ranking users according to $g(v)$ (Equation (4)), but it has some limitations. It is unable to capture certain structural patterns, such as reciprocity and communities. It also becomes unreliable as the number of followers decreases. Herein, we illustrate how the feature-based approach and the propagation approach can be incorporated with the proposed carefulness.

For the feature-based approach, we review a set of features proposed in previous works for spammer detection and then describe how they are adjusted using the carefulness. For the propagation approach, we consider a state-of-the-art algorithm, SybilBelief [Gong et al. 2014a], which has been shown to outperform other algorithms, such as SybilLimit [Yu et al. 2008], SybilInfer [Danezis and Mittal 2009], SybilRank [Cao et al. 2012], and CIA [Yang et al. 2012]. SybilBelief models the similarity between users with link weights. We discuss how the weights are derived from the carefulness. We refer to the two versions of models as the *original* and the *adjusted*, respectively. Note that the original version of features for the feature-based approach is also used in learning the carefulness.

5.1. Feature-Based Approach

5.1.1. Degrees. The first set of features includes the number of followees $|N_I(v)|$, the number of followers $|N_O(v)|$, and the number of reciprocal relations $|N_R(v)|$. An aggressive spammer follows a large number of users, but few users will follow back. Huang et al. [2013] proposed the *response rate* as the fraction of users who replied out of all recipients, and it was shown to be effective to filter aggressive spammers in an email network. In a microblogging Web site, we define the response rate as $|N_R(v)|/|N_O(v)|$ analogously. As a user tends to follow legitimate users, the response rate of a spammer is expected to be low. We define another similar feature, the *follow-back rate*, as $|N_R(v)|/|N_I(v)|$. It measures how likely a user would follow back someone who follows her.

Adjustment. However, these features can be easily manipulated by creating fake accounts and reciprocal relations between them. In general, we try to penalize links from careless and malicious users. As the features are calculated by counting links, a spammer would have less counting, making her distinct from legitimate users. Recall that the carefulness $f(u)$ is defined as the probability of identifying an actual spammer or legitimate user. A value of $1/2$ indicates that links from u are formed at random. In this case, the links should be considered noise and discarded from the graph. When $f(u) < 1/2$, the links are likely to be manipulated, so negative weights should be assigned to penalize the manipulation. For this purpose, we rescale $f(u)$ as $2f(u) - 1$ in the range $(-1, 1)$. A value of 0 corresponds to $f(u) = 1/2$.

We define the adjusted response rate as $\sum_{u \in N_R(v)} \left(2f(u) - 1\right) / |N_O(v)|$ to penalize fake links. A malicious user who follows back cannot help to manipulate this feature now. A legitimate user gets a much higher response rate because she is favored by careful users. Similarly, we adjust other degree features as sums of the rescaled carefulness.

5.1.2. Clustering Coefficients. Boykin and Roychowdhury [2005] suggested that the clustering coefficient, which measures how closely a user's friends are connected, can be used to filter email spammers. Given a node set V', we denote $E_R(V')$ as the set of reciprocal relations in the subgraph induced from V':

$$E_R(V') = \{(u, v) | u \in N_R(v) \wedge (u, v) \in V' \times V'\}. \tag{9}$$

In the context of microblogging, we propose two versions of clustering coefficients as the fraction of actual links among all possible links in different scopes:

$$C_O(u) = \frac{1}{2}|E_R(N_O(u))| / \binom{|N_O(u)|}{2}, \tag{10}$$

$$C_R(u) = \frac{1}{2}|E_R(N_R(u))| / \binom{|N_R(u)|}{2}. \tag{11}$$

Social networks are formed by communities that are tightly connected internally. A legitimate user belongs to one or more communities, so her clustering coefficient is generally high. The main difference between the two definitions lies on the scope of neighborhood under consideration. $C_O(u)$ covers the communities that the user u is attempting to join (the community members may not follow back), whereas $C_R(u)$ is limited to communities that u has reciprocal social ties with. Given a legitimate user u, $C_O(u)$ tends to be less than $C_R(u)$, as the user may follow several irrelevant communities at the same time. For a spammer u, $C_O(u)$ covers the full range of users that are annoyed. As spammers are trying to gain attentions aggressively but seldom get a follow back, $C_O(u)$ tends to be very small. Due to the different characteristics of $C_O(u)$ and $C_R(u)$, we use both of them in the detection.

Adjustment. Similarly to degrees, spammers can also manipulate clustering coefficients by linking their accounts to form fake communities. Recall that $C_O(u)$ and $C_R(u)$ count the number of reciprocal relations in a neighborhood. We adjust them by counting only "real" links as

$$C'_O(u) = \sum_{(v,w)\in E_R(N_O(u))} \frac{1}{2}\left(f(v) + f(w) - 1\right) / \binom{|N_O(u)|}{2}, \tag{12}$$

$$C'_R(u) = \sum_{(v,w)\in E_R(N_R(u))} \frac{1}{2}\left(f(v) + f(w) - 1\right) / \binom{|N_R(u)|}{2}, \tag{13}$$

where $f(v) + f(w) - 1$ is the average of the rescaled carefulness. The preceding adjustment makes the clustering coefficients of spammers even lower than those of legitimate users. In particular, if a spammer manages to make a few dense fake communities, the adjusted clustering coefficients are still low because the carefulness of the members is expected to be low.

5.1.3. PageRank. PageRank and its variants are widely used in ranking Web pages. In recent works [Chirita et al. 2005; Cao et al. 2012; Huang et al. 2013; Xue et al. 2013], it has been adapted to detect spammers in social networks. Initially, every node is assigned with the same score $1/|V|$. In each iteration, the score of a node is propagated uniformly to outgoing nodes with a damping factor d:

$$PR(v) = \frac{1-d}{|V|} + d \cdot \sum_{u\in N_I(v)} \frac{PR(u)}{|N_O(u)|}. \tag{14}$$

The key intuition for utilizing PageRank is that legitimate users rarely response to spammers, making a "cut" between the two groups. Consider a random walk on the directed graph G. At each time tick, we pick an arbitrary outgoing node of the current node as the destination with probability d, or we restart the process and pick the starting node uniformly in the entire graph with probability $1 - d$. PageRank is essentially the probability of arriving at a particular node. If we start a random walk from an arbitrary node, we are highly likely to arrive at a legitimate user eventually.

In other words, the PageRank score of a legitimate user is expected to be higher than those of spammers.

An important variant of PageRank is the reversed PageRank. Directions of links are reversed, and PageRank scores are calculated on the "reversed" graph. Ghosh et al. [2012] and Huang et al. [2013] applied the reversed PageRank to discover link farmers and spammers in social networks. As spammers create massive out-links, they receive more reversed PageRank scores from their followees, especially when they follow other spammers. Formally, it is defined as

$$RPR(u) = \frac{1-d}{|V|} + d \cdot \sum_{v \in N_O(u)} \frac{RPR(v)}{|N_I(v)|}. \tag{15}$$

Adjustment. One drawback of PageRank is that a spammer can still get a high score if she manages to attract a few top users. We fix this by introducing the personalized damping factor. For a careful user, we shall walk toward her followees with a high probability, as she knows them to be legitimate with a high confidence. However, we would want to restart the random walk to prevent the score being propagated from a malicious user. We make such adjustments by replacing the damping factor d with the carefulness $f(u)$ as

$$PR'(v) = 1 - \frac{\sum_{u \in V} PR'(u) f(u)}{|V|} + \sum_{u \in N_I(v)} \frac{PR'(u) f(u)}{|N_O(u)|}. \tag{16}$$

When a node u is arrived at, the random walk follows links starting from u with a probability of $f(u)$, or restart with a probability of $1 - f(u)$. The adjusted PageRank is calculated as the probability of arriving at a particular node in this configuration.

The reversed PageRank is less discriminating if spammers limit the number of out-links cautiously. In addition, the reversed PageRank score of a legitimate user could be higher than expected if the user happens to follow spammers. Recall that we predict if a user v is a spammer with the function $g(v)$ (Section 4.3). We define the adjusted reversed PageRank as

$$RPR'(u) = (1-d) \cdot \frac{g(u)}{\sum_{v \in V} g(v)} + d \cdot \sum_{v \in N_O(u)} \frac{RPR'(v)}{|N_I(v)|}. \tag{17}$$

The adjusted reversed PageRank is essentially a mixture of the original reversed PageRank and the prediction $g(v)$. We try to fix wrong prediction by biasing the random walk with $g(v)$. Spammers still get high scores even if they limit the number of out-links.

5.1.4. Classification. Detection of spammers is modeled as a binary classification problem. Using the adjusted features, we train a classifier with known spammers and legitimate users in a supervised approach. In addition, we would expect the classifier to estimate the probability for every user to be a spammer so that a ranking can be produced.

5.2. Propagation Approach

A recently proposed algorithm, SybilBelief [Gong et al. 2014a], extends previous propagation approaches by considering the social tie strength of adjacent users. We give a brief introduction of the model here. SybilBelief is based on a Markov random field (MRF) in undirected graphs. We transform the original directed graphs into undirected ones by keeping only reciprocal links. The node potential $\phi_v(Y_v)$ for the node v is defined as

$$\phi_v(Y_v) = \begin{cases} \theta_v & Y_v = 0 \\ 1 - \theta_v & Y_v = 1 \end{cases}, \tag{18}$$

and the edge potential $\varphi_{uv}(Y_u, Y_v)$ of the edge (u, v) is defined as

$$\varphi_{uv}(Y_u, Y_v) = \begin{cases} w_{uv} & Y_u = Y_v \\ 1 - w_{uv} & Y_u \neq Y_v \end{cases}. \tag{19}$$

The node potential $\phi_v(Y_v)$ encodes prior knowledge about v. Setting $\theta_v > 1/2$ means that v is legitimate. Setting $\theta_v < 1/2$ indicates that v is a spammer. If there is no prior knowledge, one may set $\theta_v = 1/2$. This provides the mechanism through which the training data can be incorporated in the model.

The edge potential $\varphi_{uv}(Y_u, Y_v)$ encodes the coupling strength of two adjacent nodes, u and v. A larger value of w_{uv} indicates that u and v are more similar. Specifically, $w_{uv} > 1/2$ means that the two nodes tend to be both spammers or legitimate users; $w_{uv} < 1/2$ indicates that u and v are different, and $w_{uv} = 1/2$ suggests no coupling. In the original SybilBelief algorithm, where homophily is assumed and no prior knowledge of links is known, the parameter w_{uv} is simply set to a constant of 0.9 for each $(u, v) \in E$.

Given the set of known users D, the detection is performed by finding a set of labels Y that maximizes the joint probability

$$P(Y) = \frac{1}{Z} \prod_{v \in V} \phi_v(Y_v) \prod_{(u,v) \in E} \varphi_{uv}(Y_u, Y_v). \tag{20}$$

The partition function $Z = \sum_Y \prod_{v \in V} \phi_v(Y_v) \prod_{(u,v) \in E} \varphi_{uv}(Y_u, Y_v)$ normalizes the probabilities. The inferred probability $P(Y_v = 1)$ is used to rank nodes from spammers to legitimate users.

Adjustment. We estimate the coupling strength w_{uv} with the proposed carefulness. Given that u and v follow each other reciprocally, they both have knowledge on if the other user is a spammer. This knowledge is encoded in the carefulness. A straightforward approach is to assign w_{uv} with the probability $P(Y_u = Y_v)$. As we saw in Section 4.2, it evolves the probability $r(u, v)$, which is complicated to calculate. To avoid this issue, we use the following heuristic. We consider the conditional probability $q_{uv} = P(Y_u = Y_v | u \in N_R(v))$. As defined in Section 4.2, we can derive

$$\begin{aligned} q_{uv} &= \sum_{y \in \{0,1\}} P(Y_u = y, Y_v = y | u \in N_R(v)) \\ &= \sum_{y \in \{0,1\}} P(Y_u = y | u \in N_R(v)) P(Y_v = y | u \in N_R(v)) \\ &= \sum_{y \in \{0,1\}} P(Y_u = y | (v, u) \in E) P(Y_v = y | (u, v) \in E), \end{aligned} \tag{21}$$

which can be calculated from $f(u)$ and $f(v)$ according to Equation (2). Note that according to our model stated in Equation (1), Y_u and Y_v are independent variables. Y_u and (u, v) are irrelevant, and so are Y_v and (v, u).

Due to the different nature of w_{uv} and q_{uv}, we cannot simply replace w_{uv} with q_{uv} in the edge potential $\varphi_{uv}(Y_u, Y_v)$. We estimate w_{uv} with the heuristic $w_{uv} = (1+\exp\{-\alpha q_{uv} - \beta\})^{-1}$, which ensures that w_{uv} ranges from 0 to 1; α and β are parameters that need to be determined in advance. In our initial experiments, we find that $\alpha = 1$ and $\beta = 0.8$ work well in most cases.

In summary, we first learn the carefulness according to Section 4. We then derive the probability q_{uv} from the carefulness and estimate the coupling strength w_{uv} with it. The adjusted SybilBelief algorithm is executed with the new coupling strength.

6. DATASET AND OBSERVATION

Before presenting the experimental results, we first describe our datasets from Sina Weibo and Twitter. We then analyze the behavior of spammers on the two networks. An interesting difference across networks is observed, which explains the different performance of the evaluated detection algorithms.

6.1. Dataset

6.1.1. Weibo. Sina Weibo[2] is one of the most popular microblogging Web sites in China. We crawled this dataset in May 2014 using the API of Sina Weibo. We applied the following strategy to obtain a reasonably “good” sample [Leskovec and Faloutsos 2006] from the whole Web site. We first sampled several tweets posted during April and May 2014 from the public timeline of the Web site, expecting to collect a uniform sample of active users. We ended up with 49,719 unique users as seeds. We crawled their following lists and the following lists of their followees—that is, we crawled the two-hop neighborhoods of the seed users. We did not crawl the followers, because the following lists of given users actually fully covered their relationships. Finally, we obtained a social graph containing 3.5 million nodes and 652 million directed links, among which 83 million pairs of users follow each other reciprocally. The social graph is connected, except for a few dozen isolated nodes.

In previous works, various criteria were used to identify spammers for ground truth, such as suspended accounts [Hu et al. 2013], unrelated tweets and hashtags [Benevenuto et al. 2010], social honeypots [Lee et al. 2010; Stringhini et al. 2010], and malicious URLs [Yang et al. 2012]. These criteria may be biased to certain types of spammers. We intended to cover a full range of spammers, so we decided to identify spammers manually.

We inspected profiles, tweets, and photos for spamming or normal activities. Users suspended by Sina Weibo were also included as spammers. A conservative strategy was applied in the inspection. A user was marked as spammer if only evidence of spamming activity was found. If conflicting evidence was observed (e.g., the user posted malicious tweets sometimes but interacted normally with friends at other times), we still considered the user as legitimate. If neither evidence was observed, we marked the user as unknown. This was usually due to the lack of activities—for example, only a few tweets without actual content were posted.

During the inspection, we spotted (but were not limited to) several typical patterns of spammers. A significant number of spammers post snippets from online news or blog posts, possibly trying to avoid content-based detection. We consider such users as spammers because they occasionally post URLs to malicious Web sites or online shops irrelevant to their tweets. This behavior is quite different from (legitimate) regular marketers, whose tweets are mostly relevant to their products. Some other spammers go further by copying tweets and photos from other users, including very personal ones (e.g., “my cat is sick”), making them more similar to real users. Such activities are identified by searching for those tweets and photos on the Web site and comparing timestamps of tweets and watermarks in photos. We also noticed fake accounts for the purpose of cheating in sweepstakes. Sweepstakes are used by many companies to draw attention to their products. Anyone who retweets a promotion tweet could win a prize draw. To increase the chance of winning, a spammer creates several fake accounts and

[2] http://www.weibo.com.

Table II. Statistics of Followers, Followees, and Reciprocal Relations for Spammers and a Random Sample of Users

		Weibo		Twitter	
		Mean	Median	Mean	Median
Followers (#)	Spammer	99	33	**254**	**37**
	Random	**185**	**44**	56	10
Followees (#)	Spammer	**192**	112	**450**	**211**
	Random	185	**116**	56	11
Reciprocal Relations (#)	Spammer	27	7	**153**	**16**
	Random	**47**	**18**	25	3

Note: Twitter spammers have significantly larger values of these statistics than random in contrast to Weibo spammers. Numbers in bold represent larger means or medians.

retweets from multiple promotion campaigns. We consider such users as spammers because they retweet in bulk and do not actually help the promotion. In addition, Yu et al. [2012] discovered that spammers artificially inflate top trends in Sina Weibo by retweeting from particular users in bulk. We also found such spammers in our dataset.

We must emphasize that not all spammers follow the preceding patterns. Many spammers require human comprehension to identify them. Finally, in a uniform sample of 2,000 users, we managed to identify 482 spammers and 1,432 legitimate users, leaving 86 users as unknown. The number of spammers appears to be large, but it is not surprising. As shown by Yu et al. [2012], a large fraction of trends in Sina Weibo are actually artificially inflated by fake accounts.

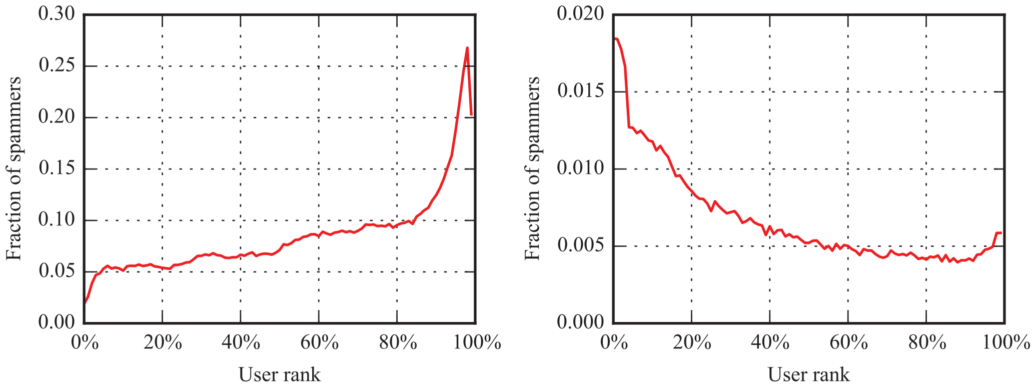
6.1.2. Twitter. We use a Twitter graph collected by Kwak et al. [2010] in September 2009. The original dataset consists of 41.7 million nodes and 1.47 billion directed links. For our study, we extracted the largest connected component, consisting of 21.3 million nodes and 1.18 billion directed links. For ground truth, we obtained a list of 145,156 suspended accounts from Gao et al. [2015]. These accounts were suspended for being spamming or just plain fake, according to Twitter's policy,[3] so we consider them as spammers in this study. We also attempted to inspect accounts manually as we did on Sina Weibo. However, Twitter is an international microblogging platform. We find it difficult for us to deeply comprehend tweets and profiles from various nations and cultures. Our analysis in the rest of this section shows that the suspended accounts share similar characteristics with manually labeled spammers in Yang et al. [2012], so we believe that our dataset is representative for this study.

6.2. Characterizing Spammers

We conduct an analysis to understand the network structure of spammers. In the Weibo dataset, we find that the inspected 2,000 users are almost disconnected, making it impossible to analyze the interaction between spammers. In this section, we consider the list of suspended users among the 3.5 million users, which can be massively crawled. We found that 297,537 users were suspended by the time of crawling. According to the Web site's policy, accounts are suspended mainly due to abusive activities, such as spamming, scam, and phishing, so they represent a subset of spammers.

6.2.1. Degrees. As the first step, we consider simple metrics characterizing spammers, such as the number of followers, the number followees, and the number of reciprocal relations. Table II shows the mean and the median of these metrics on Weibo and Twitter. For comparison, similar statistics for random samples of users are also shown here.

[3] https://support.twitter.com/articles/15790.



(a) Weibo (b) Twitter

Fig. 3. Fraction of spammers as a function of user rank. The user rank is obtained in descending order of PageRank.

We find that Twitter spammers have significantly more followers, followees, and reciprocal relations than random. Ghosh et. al [2012] explained this behavior as a result of link farming. On many Web sites, some graph-based metrics (e.g., the number of followers and PageRank) are used to estimate the importance of users. Spammers try to accumulate social capital by acquiring links massively and thereby increase the chance of spreading their spams.

However, Weibo spammers have fewer followers and reciprocal relations than random. A possible explanation is that Weibo spammers have a different spamming strategy. Yu et. al [2012] find that the trending keywords of Sina Weibo are heavily manipulated by spammers. Spammers retweet specified tweets of their customers with a large number fake accounts, expecting to increase the popularity of certain keywords. As increasing social capital is not their first goal, it makes sense for them to maintain a small number of links to avoid being noticed.

6.2.2. PageRank. Now we proceed to complicated metrics. PageRank has been widely used to rank Web pages. In recent years, it has been adapted to rank users in online social networks [Weng et al. 2010]. We are interested in how spammers are distributed according to PageRank. We rank users in descending order of their PageRank scores. We consider the fraction of spammers as a function of user rank. Figure 3 shows very distinct distributions of spammers on the two networks. A significant number of Twitter spammers succeed in boosting their ranks. However, Weibo spammers have a relatively low rank, as they have different spamming strategies.

6.2.3. Community. Previous works observe a different community structure of spammers. On Twitter, spammers tend to form tightly connected communities [Yang et al. 2012]. On Renren, a Chinese social network, spammers do not form such communities but integrate into the normal social graph [Yang et al. 2011; Zhu et al. 2012]. We observed a similar difference in our dataset. We use three metrics to quantify the interaction between spammers: graph density, reciprocity, and average distance. For each of the Weibo and Twitter networks, we calculate the metrics on the subgraph induced from spammers and the original graph, respectively (Table III). The three metrics are as follows:

—*Graph density*. Graph density is the fraction of actual links out of all possible links between nodes. In a directed graph, it is calculated as $\frac{|E|}{|V|(|V|-1)}$. It measures how

Table III. Statistics of Graph Density, Reciprocity, and Average Distance for Spammers and All Users

		Weibo	**Twitter**
Graph Density	Spammer	5.133×10^{-5}	1.790×10^{-4}
	All	5.214×10^{-5}	2.606×10^{-6}
Reciprocity	Spammer	0.124	0.329
	All	0.146	0.259
Average Distance	Spammer	3.974	3.505
	All	3.920	4.226

Note: Twitter spammers tend to form tightly connected communities, whereas Weibo spammers do not.

tightly the nodes are connected in a graph. The graph density of the original Weibo graph is 5.214×10^{-5}. The subgraph induced from spammers has a density of 5.133×10^{-5}. Graph density of the two graphs does not have significant difference, indicating that Weibo spammers appear to spread in the whole network. On Twitter, graph density of the spammer subgraph is 1.790×10^{-4}, which is greater than graph density of the original graph (2.606×10^{-6}) by two orders of magnitude. This suggests that Twitter spammers are more tightly connected than normal users.

—*Reciprocity*. We define reciprocity as the fraction of reciprocally connected user pairs out of all connected user pairs. A higher value indicates that users are more likely to follow each reciprocally. On Weibo, we find that the reciprocity of spammers (0.124) is slightly lower than the reciprocity of all users (0.146), indicating that Weibo spammers do not follow each other intentionally. However, on Twitter, the reciprocity of spammers (0.329) is significantly higher than the reciprocity of all users (0.259). This shows that Twitter spammers are trying to form tightly connected communities.

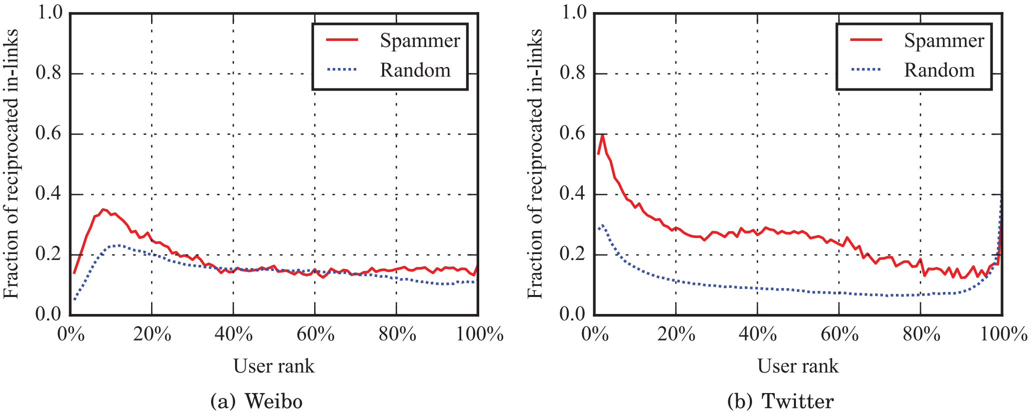
—*Average distance*. Average distance is defined as the average length of shortest paths between every pair of nodes. A lower value indicates that two nodes are more likely to be reachable via a few steps. Again, we do not observe significant difference of this metrics between spammers (3.974) and all users (3.920) on Weibo. On Twitter, the average distance between spammers (3.505) is lower than the average distance between all users (4.226). This result shows again Twitter spammers are tightly connected to each other, whereas Weibo spammers do not form such communities.

The preceding findings have the following impacts on detection. Graph-based features have very distinct distributions in the two networks. A feature that can tell spammers from legitimate users may not work in another network. Consequently, the feature-based approach needs to be carefully tuned for each network. The propagation approach (e.g., SybilBelief) assumes that spammers are tightly connected to each other. This assumption holds on Twitter but is not true on Weibo, suggesting that the propagation approach is not applicable there.

6.3. Characterizing Neighbors of Spammers

Now we proceed to the spammer's one-hop neighbors. A *spam follower* is a user who follows at least one spammer. A *spammer target* is someone who is followed by at least one spammer. It is commonly agreed that legitimate users favor only other legitimate users and avoid following spammers, but previous works have spotted several types of legitimate users who are quite likely to follow spammers, such as social capitalists [Ghosh et al. 2012] and dummies [Yang et al. 2012]. Whether this is universal in the two networks is of interest.

We find that 69.4% of Weibo users and 23.4% of Twitter users are spam followers, whereas 83.9% of Weibo users and 56.2% of Twitter users are spammer targets. This means that spamming is conducted at large scale on the two Web sites. Note that this



(a) Weibo

(b) Twitter

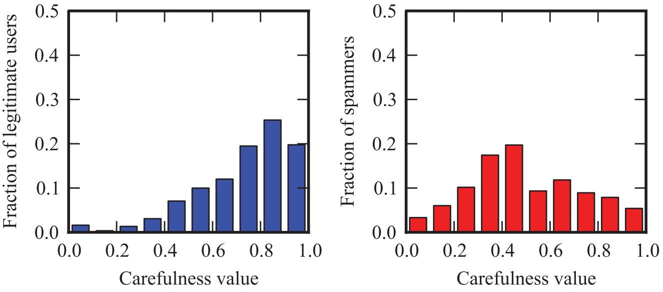
Fig. 4. Fraction of reciprocated in-links from spammers as a function of user rank. The user rank is obtained in descending order of PageRank.

result underestimates the scale of spamming, as our dataset only contains a fraction of actual spammers.

We then shift our focus on spam followers. We measure how likely a user would follow spammers by the fraction of reciprocated in-links from spammers. A value of 1 indicates that the user always follows back spammers, and a value of 0 indicates that the user never follows a spammer. On average, Weibo spammers follow back 18.3% of other spammers who follow them. This number is a bit higher than that of a random spam follower, who follows back 15% of spammers. The distinction between random spam followers and spammers is more obvious on Twitter. Twitter spammers follows back 25.5% of other spammers, whereas a random spam follower follows back only 10.7% spammers. This agrees with our previous observation that Twitter spammers are tightly connected to each other.

We further plot the fraction of reciprocated in-links from spammers as a function of user rank in Figure 4. Both Weibo and Twitter have a tendency that top users with high ranks are more likely to follow spammers. Ghosh et al. [2012] explain this behavior as a consequence of link farming, where top users encourage others to follow them by following back anyone, aiming to increase their social capital. However, a subtle distinction between the top spam followers is observed. The very top Weibo users are unlikely to follow back spammers. For example, the top 1% of Weibo users follow back only 5% of spammers who target them, which is significantly lower than average (15%). On Twitter, a similar behavior is observed for the top few users. The top 1,000 users (<0.01%) reciprocate 5.9% in-links from spammers, which is lower than average (10.7%). This suggests that the very top users do not obtain popularity via link farming.

In summary, the preceding observations show quite different characteristics of spammers on Weibo and Twitter. Twitter spammers have larger degrees, and quite a few of them succeed in gaining high ranks. However, Weibo spammers tend to be modest by maintaining relatively few degrees, and most of them have low ranks. Another distinction is that Twitter spammers tend to form tightly connected communities, whereas Weibo spammers spread and hide in the whole network. The two distinct behaviors have both been observed in previous works in different networks [Yang et al. 2011, 2012; Ghosh et al. 2012]. Whether the proposed carefulness can be leveraged to detect spammers with such distinct behaviors is of interest.



(a) Legitimate users (b) Spammers

Fig. 5. Distributions of the carefulness for legitimate users and spammers (Weibo).

7. EXPERIMENTS

In this section, we present our experimental results. Our first concern is how the learned function $f(u)$ reflects the carefulness of users, so we conduct an empirical study with various side information for justification. We then evaluate the performance of spammer detection that is aided by $f(u)$.

7.1. Carefulness

As the first step, we calculated the carefulness as described in Section 4 for all users. We compared the result with various side information to validate our method.

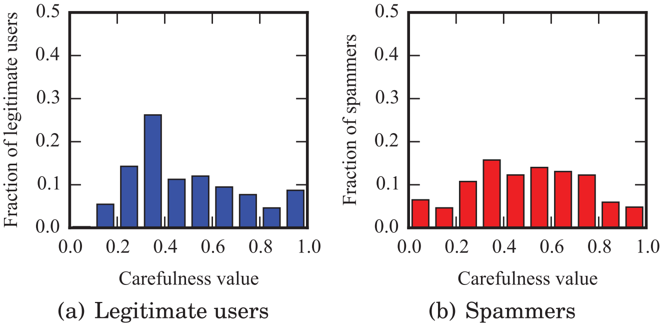
7.1.1. Spammers. We first studied the difference between legitimate users and spammers in terms of the carefulness. We grouped legitimate users and spammers based on $f(u)$ and computed the fraction of users in each group.

On Weibo, the result (Figure 5) shows an obvious tendency of high value for legitimate users and low value for spammers, whose averages are 0.730 and 0.497, respectively. Legitimate users are quite careful in avoiding spammers (e.g., $f(u) > 0.5$ for 87% of legitimate users). We also find that most legitimate users concentrate in the range $[0.8, 0.9]$, but relatively few of them are extremely careful (e.g., $f(u) > 0.9$ for 20% of legitimate users). This is consistent with our observation in Section 4.1 that a large fraction of legitimate users follow at least one spammer.

However, spammers have various carefulness values in all ranges. Most spammers appear to be careless, and the others are either malicious or careful. This could be explained by the strategies of spammers to seek user IDs. Most spamming accounts are controlled by automated scripts, so they follow whoever they see, making them appear to be careless. Some accounts of spammers are used to boost the reputation of other spammers [Yang et al. 2012] or their (legitimate) customers [Yu et al. 2012], so they behave either maliciously or carefully.

Although the carefulness is learned from users' followees rather than themselves, the preceding results show the correlation between it and the type of users. The result is roughly consistent with the assumption in previous works that a legitimate user favors other legitimate users, but more importantly, the cases that legitimate users follow spammers are captured by our method.

However, in the Twitter dataset, there is no obvious distinction of the distributions of carefulness (Figure 6). The average carefulness values of legitimate users and spammers are 0.494 and 0.496, respectively. An explanation is that suspended spammers only consist of a fraction of actual spammers on Twitter. Many other spammers are mistakenly regarded as legitimate. The followers of these spammers are thus inferred



(a) Legitimate users (b) Spammers

Fig. 6. Distributions of the carefulness for legitimate users and spammers (Twitter).

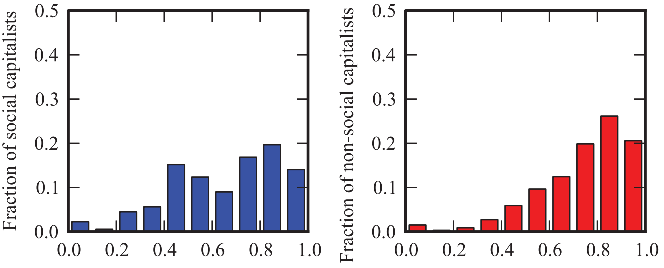
as careful, although many of the followers are also spammers. This in turn confuses the carefulness model.

To validate this explanation, we conducted an experiment using the Weibo dataset. Among the 2,000 labeled users, we take only suspended users (195 users) as spammers and regard other users as legitimate (including those 287 manually labeled spammers). This simulates the scenario in the Twitter dataset. We train the carefulness model with the modified Weibo dataset. The result shows that the distributions of carefulness for legitimate users and spammers are similar, which is similar to what we observed in the Twitter dataset. Although the carefulness of individual users may be inferred incorrectly due to the incomplete labels in the Twitter dataset, we find that it is still helpful for the detection when the values are aggregated (Section 7.2).

7.1.2. Social Capitalists. As shown in Ghosh et al. [2012], social capitalists are trying to increase their social capital by following back anyone who follows them, so it is reasonable to assume that the carefulness of social capitalists is around 1/2. In the Weibo dataset, we identified social capitalists from known legitimate users as follows. We obtained the category of a user, such as civilians, famous artists, or enterprises, via the API of Sina Weibo. We considered a user as a social capitalist if she was not a civilian. For users missing such information, we inspected them manually. Generally, we found that most social capitalists were trying to promote their tweets and gain attention from others, whereas non–social capitalists simply subscribe to popular accounts and communicate with friends. We ended up with 12.4% of legitimate users as social capitalists. We did not identify social capitalists on Twitter, so we only present results of the Weibo dataset.

The distribution of social capitalists (Figure 7(a)) shows two distinct peaks. In the first peak, 28% of social capitalists are in range [0.4, 0.6], indicating a careless behavior of them. This is expected by the definition of social capitalists. In the second peak, 34% of social capitalists have carefulness values greater than 0.8. We inspected social capitalists with top carefulness values and found that they are mainly popular bloggers or government and related organizations. A popular blogger typically has hundreds of thousands of followers but only dozens of followees. Their tweets are related to popular topics like health care, jokes, and lifestyle. We are unclear about how they gain so many followers, but apparently it is not via following every follower, so it makes sense to consider them as careful. For government and related organizations, they do not actually need to apply such strategies because they are known as authoritative by everybody.

Most non–social capitalists are inferred as careful, because they use microblogs as a regular social network service rather than a platform for promoting. Note that



(a) Social capitalists (b) Non-social capitalists

Fig. 7. Distributions of the carefulness for social capitalists and non–social capitalists (Weibo).

Table IV. Average Carefulness of Users Grouped by Binary (Yes/No) Profile Features

	Weibo		Twitter	
Feature	Yes	No	Yes	No
Posted any tweet?	**0.700**	0.682	**0.509**	0.386
Tweet in favorite?	**0.750**	0.603	**0.540**	0.453
Nonempty bio?	**0.728**	0.638	**0.542**	0.485
Custom domain?	**0.735**	0.699	N/A	N/A
Custom profile image?	**0.701**	0.602	**0.538**	0.434
Personal Web site?	**0.717**	0.696	0.**546**	0.478
Direct message from stranger?	0.663	**0.704**	N/A	N/A
Comment from stranger?	0.696	**0.733**	N/A	N/A
Public tweets?	N/A	N/A	0.480	**0.580**
Geolocation enabled?	0.697	**0.740**	**0.531**	0.485
Is verified?	**0.700**	0.549	**0.766**	0.485

Note: Not all features are available on both Weibo and Twitter. Numbers in bold represent larger values of carefulness.

Figure 7(b) is expected to be similar to Figure 5(a) because the majority of legitimate users are non–social capitalists.

7.1.3. Profiles. We also crawled the profiles of users on both Weibo and Twitter. We extracted three groups of binary features from the profiles, focusing on inactive users, privacy settings, and user verification. For each feature, we split the users into two groups according to the feature value (yes/no) and calculate the average of their carefulness $f(u)$. The result (Table IV) shows that active users (e.g., those who have ever posted a tweet, saved a tweet in favorite, or written a bio) are more careful. Active users learn about spammer's strategies while browsing the Web site, so they are better at avoiding spammers. We also find that users who are more concerned about privacy (e.g., disallowing direct messages or comments from strangers) are inferred as more careful. This is reasonable because these users are not likely to follow others at random, as otherwise their privacy will breached. Verified users (including individuals and organizations) are much more careful than ordinary users. Verified users are required to expose their real identities in the Web site, so they should maintain their accounts seriously. In summary, although the carefulness is learned based on graph structure, these results show consistent correlation with profiles on both Weibo and Twitter.

7.2. Detection

Now we evaluate the performance of the two detection approaches described in Section 5. For the feature-based approach, we start with individual features and then combine the features to train a classifier. For the propagation approach, we compare the original SybilBelief algorithm and its adjusted version.

Table V. AUC of Detection with Individual Features

	Weibo			Twitter		
Feature	Original	Adjusted	Gain	Original	Adjusted	Gain
Number of followees	0.564	**0.721**	28%	**0.841**	0.537	−36%
Number of followers	0.600	**0.810**	35%	0.728	**0.764**	5%
Number of reciprocal relations	0.712	**0.822**	15%	**0.737**	0.732	−1%
Response rate	0.709	**0.820**	16%	**0.747**	0.710	−5%
Follow-back rate	0.640	**0.824**	29%	0.641	**0.743**	16%
Clustering coefficient $C_O(u)$	0.851	**0.861**	1%	0.651	**0.726**	12%
Clustering coefficient $C_R(u)$	0.796	**0.812**	2%	0.554	**0.695**	25%
PageRank	0.673	**0.745**	11%	0.612	**0.644**	5%
Reversed PageRank	0.635	**0.736**	16%	0.854	**0.862**	1%

Numbers in bold represent larger values of AUC.

7.2.1. Criteria. We adopt the standard notion of true-positive rate and false-positive rate to measure how successful the detection is. We regard spammers as positive samples and legitimate users as negative samples, respectively. The *true -positive rate* is defined as the fraction of correctly identified spammers out of actual spammers. The *false -positive rate* is defined as the fraction of legitimate users that are misclassified out of actual legitimate users. The trade-off between the true-positive rate and false-positive rate can be visualized by the receiver operating characteristic (ROC). We quantify the overall performance with the area under the curve (AUC).

7.2.2. Individual Features. We used each feature described in Section 5.1 alone to detect spammers and compared the performance of the original version and the adjusted version. The result (Table V) shows consistent improvements over the original ones, except for degree features in the Twitter dataset.

Degree features can be easily manipulated by spammers by connecting fake accounts. These features are expected to work poorly at the first place. On Weibo, when the features are adjusted with the carefulness, a significant improvement occurs. However, the improvement is not significant on Twitter. The number of followees turns out to be much less predictive when adjusted. As we observed in Table II, the number of followees tend to be small for spammers on Weibo but large on Twitter. When the followees are weighted with carefulness, this feature is even smaller on both Weibo and Twitter. As a consequence, it is more predictive on Weibo but less predictive on Twitter. The performance of response rate and follow-back rate is also reduced slightly after adjustment on Twitter.

Clustering coefficients are the most effective features on Weibo but not so predictive on Twitter. This is a direct consequence of the community structure of spammers. As we saw in Section 6.2, Twitter spammers form tightly connected communities, whereas Weibo spammers spread in the whole network. When the clustering coefficients are adjusted by carefulness, some “fake” communities formed by spammers are discarded, resulting in significant improvement on Twitter. On Weibo, the improvement is quite slight.

PageRank features perform equally well on Weibo, even after the adjustment. On Twitter, PageRank is the least predictive feature. However, its variant, the reversed PageRank, turns out to be the most effective one. An explanation is that some Twitter spammers focus on boosting their ranks to better spread spams [Ghosh et al. 2012]. These spammers are mistaken with other (legitimate) influential users. Somehow they failed to manipulate their reverse PageRank. However, Weibo spammers have a different spamming strategy [Yu et al. 2012], and ranking is not their main objective.

7.2.3. Evaluation. For the feature-based approach, we combined all adjusted features and used random forests to perform the detection (`RF-adjusted`). In our initial experiments, we tried several popular classifiers, including logistic regression, support vector machines with different kernels, and other types of decision trees. It turned out that random forests outperformed others significantly in terms of AUC. For the propagation approach, we considered the SybilBelief algorithm that is adjusted with carefulness (`SybilBelief-adjusted`). For comparisons, we employed the following methods as baselines:

- —We use the original version of features to train another random forest (`RF-original`) to examine the effect of the carefulness.
- —We consider the original SybilBelief algorithm where the parameter w_{uv} is set to 0.9 (`SybilBelief-original`).
- —In Section 4.3, we estimate the label of a user with the function $g(v)$ (see Equation (4)) and optimize it directly. We take it as a baseline to compare with.
- —TrustRank [Gyöngyi et al. 2004] was proposed to detect Web spams, and we adapt it for spammer detection in a microblogging site. TrustRank requires a few known good nodes to start the propagation. The seed nodes are crucial to a successful detection. We evaluated several strategies for seed selection, including high PageRank, high reversed PageRank, and uniform sampling. It turned out that uniform sampling yields the best performance, so we took a sample of 100 legitimate users as seeds.

In addition to the preceding baselines, we also considered matrix factorization-based methods recently proposed by Zhu et al. [2012] and Hu et al. [2013]. However, both methods require particular auxiliary information. The first method requires a bipartite graph that encodes user activities (e.g., visiting albums and sharing), and it is designed for undirected graphs. The second method is designed for microblogging networks but needs the content of tweets. With only the graph structure, the preceding two methods cannot work properly, so we do not compare them here.

The results (Figures 8 and 9) show that the estimated label $g(v)$ outperforms TrustRank significantly. This confirms our observation that a legitimate user does not always follow legitimate users. It is necessary to model the carefulness separately. TrustRank seems unable to work properly in the Twitter dataset. TrustRank is essentially a biased version of PageRank. Legitimate users are expected to be ranked high, but some Twitter spammers succeed in boosting their PageRank by acquiring links from legitimate users. We tried including various numbers of seeds (up to 50% of labeled users), but the AUC never exceeds 0.7.

SybilBelief and its adjusted version outperform $g(v)$ on Twitter but is less predictive on Weibo. SybilBelief is built on the assumption of homophily (i.e., connected users are similar). Homophily of spammers is observed on Twitter, as they form tightly connected communities (Section 6.2). However, Weibo spammers do not have such a strong tendency of connecting each other. This explains why SybilBelief works better on Twitter. We cast the carefulness to the coupling strength as described in Section 5.2. This adjustment improves the performance of SybilBelief on both Weibo and Twitter.

Random forests with a rich set of features outperform the preceding methods, which is expected because the proposed features capture a wide range of patterns in social networks. For example, the clustering coefficients are good measures to describe the community structure of the graph, whereas TrustRank, $g(v)$, and SybilBelief are unable to capture such patterns. By adjusting features with the proposed carefulness, the performance is further improved. As the original version of features treats every link equally, it could be manipulated by establishing fake social ties. By weighting links with the carefulness, such an effect is reduced by a considerable extent, making the features more effective for the detection.

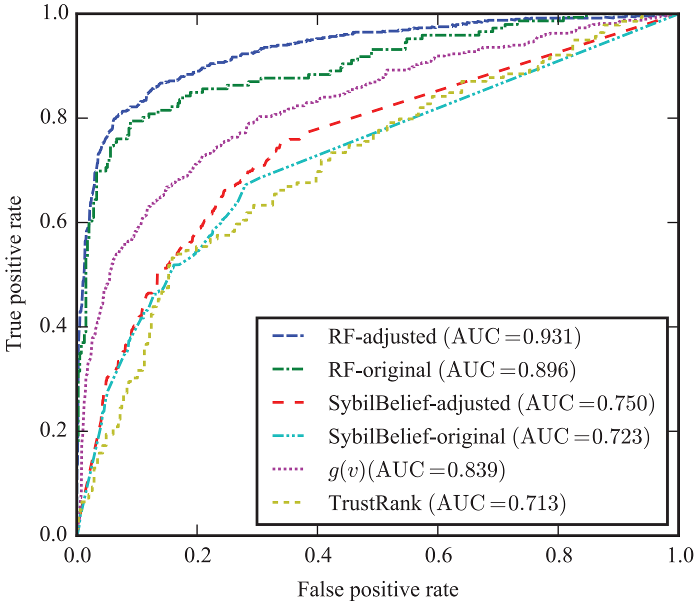


Fig. 8. ROC curves of different detection methods (Weibo).

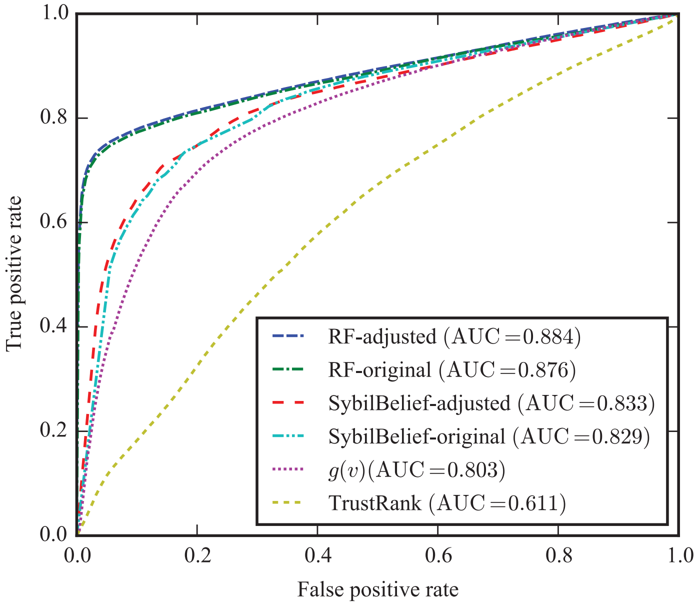


Fig. 9. ROC curves of different detection methods (Twitter).

We conducted an online test on Weibo to verify our results. We sampled two groups of spammers identified by the feature-based approach that were not suspended by the time of our manual inspection. For the first group, we reported them to Sina Weibo via the "report abuse" link in the profile page. After a week, 41% of the reported

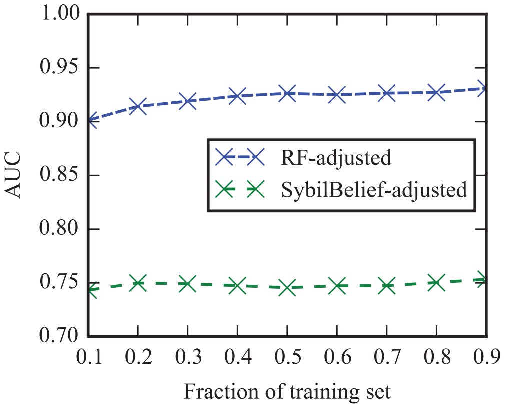


Fig. 10. AUC of detection with varying sizes of the training set for carefulness (Weibo).

spammers were suspended, whereas the others still remained active. We then reported the remaining spammers again, 27% more spammers were suspended. In total, 68% of the spammers in the first group were eventually suspended. We further examined the remaining active spammers carefully and found evidence for abusive activities (e.g., posting advertisements and suspicious URLs). As a comparison, we did not report the second group but kept monitoring them. None of them was suspended in the first month. After 7 months, only 16% of them were eventually suspended. Although we may keep reporting the first group, this result shows that those spammers were difficult for the Web site to detect currently, and our method is effective in capturing such spammers.

In summary, we find that both the feature-based approach and the propagation approach can be improved by incorporating with the proposed carefulness in a proper way. The improvement is significant on Weibo but slight on Twitter. As discussed in Section 7.1, our Twitter dataset only consists of a fraction of actual spammers. The carefulness learned from such an incomplete training set turns out to be less predictive and is thus less effective in detecting spammers. We believe that the detection performance on Twitter can be further improved if a better training set is provided.

7.2.4. Size of the Training Set. The preceding experiments show that the carefulness can be leveraged to enhance spammer detection. A set of labeled spammers and legitimate users is required to learn the carefulness. Similarly to most machine learning algorithms, the size of the training set plays an important role on the performance of the trained model. We are interested in how the size of the training set affects the carefulness. As we do not have any ground truth for the carefulness, we evaluate how the carefulness that is learned with a training set of varying sizes could affect the performance of spammer detection.

We made training sets consisting of 10%, 20%, . . . , and 90% labeled users and left the rest as testing sets. The carefulness was learned from these training sets, respectively, and incorporated with spammer detection. All experiments were performed 10 times independently, and averages were reported. The result (Figures 10 and 11) shows that increasing the size of the training set has a positive effect on the performance of detection. For both the feature-based approach and the propagation approach, no significant improvement is observed when the size of the training set goes above 50%. This suggests that a relatively small number of labeled users are sufficient to learn the carefulness.

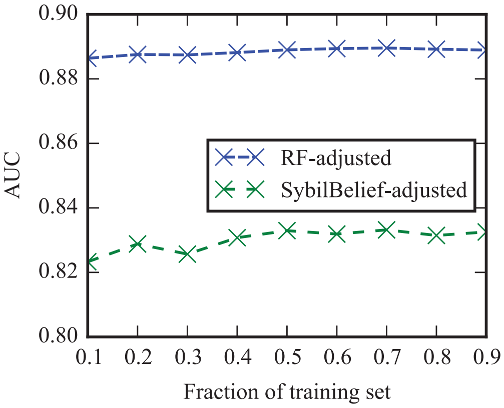


Fig. 11. AUC of detection with varying sizes of the training set for carefulness (Twitter).

8. DISCUSSION

In this section, we discuss several technical issues of our approach. We also illustrate how other graph-based applications (e.g., link prediction) can benefit from the proposed carefulness.

8.1. Loss Function

The choice of loss function plays an important role in most machine learning algorithms. As the learned carefulness is incorporated with various algorithms, it is unclear how to choose a loss function that optimizes the evaluation metrics (e.g., AUC) directly. Therefore, we evaluated several choices empirically in our initial experiments.

The first choice that we evaluated is the maximum-likelihood estimation, which seeks the value of **w** that maximizes the probability of observed spammers and legitimate users according to Equation (2). We also tried the absolute error loss (i.e. minimizing the absolute difference between $g(v)$ and the actual label). Experiments show that these approaches only provide slight improvement over the original detection methods.

Another approach is adapted from Backstrom and Leskovec [2011]. We denote D^+ and D^- as spammers and legitimate users in the training set, respectively. We require that $g(u) > g(v)$ for any $u \in D^+$ and $v \in D^-$ (i.e., a spammer should always be estimated as more suspicious than legitimate users). Although this requirement is too hard to satisfy in practice, it is relaxed with a loss function as

$$L(\mathbf{w}) = \sum_{u \in D^+, v \in D^-} h(g(v) - g(u)) + \frac{\lambda}{2} \sum_{i=1}^{k} w_i^2, \tag{22}$$

where $h(x) = (1 + \exp\{-x/b\})^{-1}$ is the Wilcoxon-Mann-Whitney (WMW) loss [Yan et al. 2003] with width b. Our evaluation shows that this loss function yields comparable performance of the squared error loss. However, the WMW loss is calculated for $|D^+||D^-|$ pairs of nodes, which may cause scalability issue when the training set becomes larger. Therefore, we use the squared error loss as the best choice in our method.

8.2. Detection without Training

Sometimes it is difficult to collect sufficient labeled data to train a classifier (e.g., restricted human inspection due to security and privacy concerns) [Zhu et al. 2012], and zero-day spammers that were never observed before [Lee et al. 2010]. Herein, we introduce a heuristic that does not need any labeled data.

We consider a simplified model of Equation (1), where $p(v) = P(Y_v = 1)$ denotes the estimated probability for a user v being a spammer, and the probability $r(u, v)$ of a "following" action is simplified as $r(u)$. The carefulness $f(\cdot)$ is no longer a function feature but only a parameter that needs to be estimated. We have the probability of creating a link (u, v) as

$$\begin{aligned} P((u,v) \in E) &= \sum_{y \in \{0,1\}} P((u,v) \in E | Y_v = y) P(Y_v = y) \\ &= \big(f(u) + p(v) - 2f(u)p(v)\big) r(u). \end{aligned} \tag{23}$$

We assume that the observed social graph G is generated based on this model; $f(\cdot)$, $p(\cdot)$, and $r(\cdot)$ are parameters that need to be inferred from the model. We fit the model to the given graph by maximizing the overall likelihood. In our experiments, ranking users with only $p(\cdot)$ yields an AUC of 0.864 on Weibo and 0.820 on Twitter, which are reasonably good compared to the unsupervised features in Table V.

8.3. Link Prediction

In a microblogging social network, the link prediction problem [Liben-Nowell and Kleinberg 2003] can be formulated as follows. Given a snapshot of the social network, is it possible for a given user to follow another one in the future?

We are interested in how the proposed carefulness can improve the performance of such applications. Many existing link prediction methods are based on the idea of closing triangles (i.e., a user connects to a friend of a friend). If two users u and v share more common friends, they are considered more similar, and thus it is more likely for them to form a connection. However, if the common friends are careless or even malicious, we would expect that the links between u (or v) and them are formed randomly. In this case, we are less confident to say that they will be connected in the future.

We consider some typical methods to predict if u will follow v and introduce how to adjust them based on thepreceding intuition. Note that we are not intended to propose new methods for link prediction here. For simplicity, we denote $f(S) = \sum_{u \in S} f(u)$ for a given node set S. The following methods are considered:

—*Common friends*. We define the number of common friends in a directed graph as $|N_O(u) \cap N_I(v)|$. We adjust this measure by weighting common friends with their carefulness as $f(N_O(u) \cap N_I(v))$.

—*Jaccard coefficient*. The Jaccard coefficient measures the similarity between two friend lists as $\frac{|N_O(u) \cap N_I(v)|}{|N_O(u) \cup N_I(v)|}$. We adjust it as $\frac{f(N_O(u) \cap N_I(v))}{f(N_O(u) \cup N_I(v))}$.

—*Adamic-Adar*. Adamic and Adar [2003] considered a related measure $\sum_{w \in N_O(u) \cap N_I(v)} \frac{1}{\log |N_R(w)|}$. When the carefulness is incorporated, it is defined as $\sum_{w \in N_O(u) \cap N_I(v)} \frac{f(w)}{\log |N_R(w)|}$.

—*Preferential attachment*. In the preferential attachment model, it is assumed that the probability of forming a new link is proportional to degrees (i.e., $|N_O(u)||N_I(v)|$). It is adjusted as $f(N_O(u)) f(N_I(v))$.

—*Random walk*. Random walk–based methods [Liben-Nowell and Kleinberg 2003; Backstrom and Leskovec 2011] have been shown to be effective for the link prediction problem. The random walk starts at u, and it returns to u with probability $1 - \alpha$ at each step. We redefine the restart probability with the carefulness similarly to Section 5.1.3. When the random walk arrives at a node w, it returns to u with probability $1 - f(w)$.

Table VI. AUC of Link Prediction

	Weibo			Twitter		
Measure	Original	Adjusted	Gain	Original	Adjusted	Gain
Common friends	0.761	**0.782**	2.8%	0.766	**0.887**	15.8%
Jaccard's coefficient	0.678	**0.688**	1.5%	0.617	**0.694**	12.5%
Adamic-Adar	0.786	**0.787**	0.1%	0.861	**0.917**	6.5%
Preferential attachment	0.563	**0.571**	1.4%	0.720	**0.736**	2.2%
Random walk	0.948	**0.964**	1.7%	0.824	**0.864**	4.9%

We focus on predicting links between nodes that are two-hops from a given node [Backstrom and Leskovec 2011]. A node pair (u, v) is considered as a positive sample if there is a link from u to v. In a practical scenario, there is no reason to predict links from or to spammers. We hereby only consider pairs of nodes in which both are known to be legitimate.

In the Weibo dataset, we find that the 2,000 labeled users mentioned in Section 6.1 are mostly apart from each other, resulting in an insufficient number of testing samples. Thus, we made another uniform sample of 100 users from our dataset and inspected them manually. For each user, 10 followees were sampled and also inspected manually. We ended up with 19,191 pairs of legitimate users for testing. For the Twitter dataset, we sampled 2,000 pairs of nodes for testing. We measure the performance of prediction by the AUC.

The result (Table VI) shows consistent improvements over the original methods. The Adamic-Adar measure estimates how serious a common friend is with the degree, which follows a similar idea of our approach. As a result, incorporating the carefulness does not bring much additional information, and the performance is similar to the original one. We have also tried including spammers in the test set. It turns out that the performance drops slightly, indicating that the carefulness is only helpful for real users.

Various additional information, such as graph attributes [Backstrom and Leskovec 2011; Gong et al. 2014b], contents [Gao et al. 2011], and locations [Wang et al. 2011], has been shown to be useful for the link prediction problem. Interestingly, spammers who are considered harmful for social networks turn out to be beneficial for the prediction in an unusual way. Generally, as users interact with spammers in social networks, certain traits are exhibited, which help us better understand the behavior of users. In our case, new links can be partially explained by the carefulness. By learning the carefulness via spammers, we can better predict new links.

9. CONCLUSION

As the behavior of users varies when they are following someone else in a microblogging Web site, we propose a framework to quantify the carefulness of a user. We develop a supervised learning algorithm to estimate the carefulness. As the carefulness is not directly visible, we conduct studies over different types of indirect evidence to justify our result. We illustrate the difference in a spammer's behavior on two popular microblogging Web sites, Sina Weibo and Twitter, and explain why a detection algorithm may not work equally well on the two networks. We then show how the robustness of detection algorithms can be enhanced using the proposed carefulness. Our experiments show that the carefulness is indeed effective for the detection.

Our observation on the Sina Weibo and Twitter datasets raises an important issue in designing spammer detection algorithms. Many existing works are built on certain assumptions of spammer's behavior. For example, spammers are tightly connected or apart from each other, so an algorithm that works well on one network may not work properly on another network, if the assumption is not true. This limits the domain of application of these algorithms. We are not intended to propose a "universal" algorithm that works on all types of networks, but we show that the proposed carefulness is helpful to improve the robustness of existing algorithms, which in turn expands the domain of application.

There are many potential future works based on this work. It would be interesting to combine the content information (e.g., tweets, photos, and profiles) to enhance the inference of carefulness. It would also be interesting to apply the proposed method to other types of social networks, such as email communication networks.

The carefulness itself could be of its own interest for research. It provides a way to understand user behavior via spammer data. It can be used as tool to analyze and interpret user behaviors in a microblogging Web site. As shown in this article, it is helpful for not only spammer detection but also link prediction. We believe that there are many other potential applications of the proposed carefulness.

The proposed model of carefulness can be extended to capture more fine-grain patterns. Similarly to most spammer detection systems, the false-positive rate and false-negative rate of a user are not necessarily the same. Although a user can recognize all legitimate users correctly, she may make mistakes about spammers. The two cases can be modeled separately (e.g., $f^+(u)$ for false positives and $f^-(u)$ for false negatives). Another possible extension is the pairwise carefulness $f(u, v)$. When a user u is about to follow a spammer v, the decision is also affected by how well v pretends to be legitimate. We leave these extensions for future work.

Our method can be seen as a passive way to utilize users' own knowledge (recognizing spammers or legitimate users) to aid spammer detection. As spammers are upgrading themselves rapidly, it is exhausting to upgrade the detection system at the same time to win the fight. We believe that users should play a central role in the campaign, as they are quick to notice new types of spam. Most users are also motivated to fight spams because spams cause financial lost and privacy leak of users. In this sense, we believe that characterizing users themselves and leveraging their power to detect spams is a promising direction toward this problem.

REFERENCES

Lada A. Adamic and Eytan Adar. 2003. Friends and neighbors on the Web. *Social Networks* 25, 3, 211–230.

Lars Backstrom and Jure Leskovec. 2011. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*. 635–644.

Fabrıcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgılio Almeida. 2010. Detecting spammers on Twitter. In *Proceedings of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse, and Spam Conference*. 12.

Yazan Boshmaf, Dionysios Logothetis, Georgos Siganos, Jorge Lería, Jose Lorenzo, Matei Ripeanu, and Konstantin Beznosov. 2015. Íntegro: Leveraging victim prediction for robust fake account detection in OSNs. In *Proceedings of the 2015 Network and Distributed System Security Symposium*.

P. O. Boykin and V. P. Roychowdhury. 2005. Leveraging social networks to fight spam. *Computer* 38, 4, 61–68.

Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. 2012. 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*. 15.

Paul-Alexandru Chirita, Jörg Diederich, and Wolfgang Nejdl. 2005. MailRank: Using ranking for spam detection. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. 373–380.

George Danezis and Prateek Mittal. 2009. SybilInfer: Detecting sybil nodes using social networks. In *Proceedings of the ISOC Network and Distributed System Security Symposium*.

Peng Gao, Neil Zhenqiang Gong, Sanjeev Kulkarni, Kurt Thomas, and Prateek Mittal. 2015. SybilFrame: A defense-in-depth framework for structure-based sybil detection. arXiv:1503.02985.

Sheng Gao, Ludovic Denoyer, and Patrick Gallinari. 2011. Temporal link prediction by integrating content and structure information. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. 1169–1174.

Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Phani Gummadi. 2012. Understanding and combating link farming in the Twitter social network. In *Proceedings of the 21st International Conference on World Wide Web*. 61–70.

Neil Zhenqiang Gong, Michael Frank, and Payal Mittal. 2014a. SybilBelief: A semi-supervised learning approach for structure-based sybil detection. *IEEE Transactions on Information Forensics and Security* 9, 6, 976–987.

Neil Zhenqiang Gong, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine Runting Shi, and Dawn Song. 2014b. Joint link prediction and attribute inference using a social-attribute network. *ACM Transactions on Intelligent Systems and Technology* 5, 2, Article No. 27.

Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. @Spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*. 27–37.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating Web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Data Bases*. 576–587.

Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. 2007. Fighting spam on social Web sites: A survey of approaches and future challenges. *IEEE Internet Computing* 11, 6, 36–45.

John Hopcroft, Tiancheng Lou, and Jie Tang. 2011. Who will follow you back? Reciprocal relationship prediction. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. 1137–1146.

Xia Hu, Jiliang Tang, and Huan Liu. 2014. Leveraging knowledge across media for spammer detection in microblogging. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 547–556.

Xia Hu, Jiliang Tang, Yanchao Zhang, and Huan Liu. 2013. Social spammer detection in microblogging. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. 2633–2639.

Junxian Huang, Yinglian Xie, Fang Yu, Qifa Ke, Martin Abadi, Eliot Gillum, and Z. Morley Mao. 2013. SocialWatch: Detection of online service abuse via large-scale social graphs. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer, and Communications Security*. 143–148.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*. 591–600.

Kyumin Lee, James Caverlee, and Steve Webb. 2010. Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 435–442.

Jure Leskovec and Christos Faloutsos. 2006. Sampling from large graphs. In *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining*. 631–636.

David Liben-Nowell and Jon Kleinberg. 2003. The link prediction problem for social networks. In *Proceedings of the 12th International Conference on Information and Knowledge Management*. 556–559.

Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2010. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*. 1–9.

Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. 2011. Suspended accounts in retrospect: An analysis of Twitter spam. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*. 243–258.

Binghui Wang, Le Zhang, and Neil Zhenqiang Gong. 2017. SybilSCAR: Sybil detection in online social networks via local rule based propagation. In *Proceedings of the IEEE International Conference on Computer Communications*.

Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. 2011. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining*. 1100–1108.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. TwitterRank: Finding topic-sensitive influential Twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. 261–270.

Q. Xu, E. W. Xiang, Q. Yang, J. Du, and J. Zhong. 2012. SMS spam detection using noncontent features. *IEEE Intelligent Systems* 27, 6, 44–51.

Jilong Xue, Zhi Yang, Xiaoyong Yang, Xiao Wang, Lijiang Chen, and Yafei Dai. 2013. VoteTrust: Leveraging friend invitation graph to defend against social network sybils. In *Proceedings of the 32nd IEEE International Conference on Computer Communications*. 2400–2408.

Lian Yan, Robert H. Dodier, Michael Mozer, and Richard H. Wolniewicz. 2003. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In *Proceedings of the 20th International Conference on Machine Learning*. 848–855.

Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. 2012. Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on Twitter. In *Proceedings of the 21st International Conference on World Wide Web*. 71–80.

Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y. Zhao, and Yafei Dai. 2011. Uncovering social network sybils in the wild. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement*. ACM, New York, NY, 259–268.

Sarita Yardi, Daniel Romero, and Grant Schoenebeck. 2009. Detecting spam in a Twitter network. *First Monday* 15, 1.

Haifeng Yu, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao. 2008. SybilLimit: A near-optimal social network defense against sybil attacks. In *Proceedings of the IEEE Symposium on Security and Privacy*. 3–17.

Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. 2006. SybilGuard: Defending against sybil attacks via social networks. *Computer Communication Review* 36, 4, 267–278.

L. L. Yu, S. Asur, and B. A. Huberman. 2012. Artificial inflation: The real story of trends and trend-setters in Sina Weibo. In *Proceedings of the International Conference on Privacy, Security, Risk, and Trust, and the International Conference on Social Computing*. 514–519.

Yin Zhu, Xiao Wang, Erheng Zhong, Nathan Nan Liu, He Li, and Qiang Yang. 2012. Discovering spammers in social networks. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*.

Received December 2015; revised May 2016; accepted March 2017