# Leveraging Careful Microblog Users for Spammer Detection

Hao Fu
University of Science and Technology of China
fuch@mail.ustc.edu.cn

Xing Xie
Microsoft Research
xingx@microsoft.com

Yong Rui
Microsoft Research
yongrui@microsoft.com

## ABSTRACT

Microblogging websites, e.g. Twitter and Sina Weibo, have become a popular platform 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. While 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 for spammers on purpose is also observed. We regard this behavior as a useful information for spammer detection. In this paper, 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 then illustrate how spammer detection can be improved in the aid of the proposed measure. Evaluation on a real dataset with millions of users and an online testing are performed on Sina Weibo. The results show that our approach indeed capture the carefulness, and it is effective to detect spammers. In addition, we find that the proposed measure is also beneficial for other applications, e.g. 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*

## Keywords

Microblog, social graph, spammer detection, user behavior

## 1. INTRODUCTION

In recent years, microblogging websites, e.g. Twitter and Sina Weibo, have gained increasing popularity. With rapidly growing influence among users, microblogging websites have become a universal platform for sharing personal experience, marketing, mass media, and public relationship. Similarly to other online social networking websites [11], spammers have discovered microblogging as an appealing platform to spread spams with fake accounts. Spams do not only annoy users but also lead to financial loss and privacy risks of users. Effective 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 spam strategies rapidly to race with the development of detection systems. A detection system that is able to capture most of the spammers in this month may fail in the next month. For example, it has long been a common practice for email server administrators to update spam filters frequently. In order to camouflage themselves, spammers may manipulate the profiles, content of tweets, and social relationship of their accounts. Tweets and profiles have been shown to be a good information source for detection [14, 19], but they can be faked by spammers if they wish. In addition, access to the content is sometimes restricted due to privacy concerns [29].

In a microblogging website, a user decides who to follow based on her own knowledge. While spammers can simulate normal patterns of links 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 paper, we focus on detecting spammers based on links.

It is intuitive and necessary for spammers to follow legitimate users, so that spams can be spread. However, conflicting observations have been made on whether spammers would connect to other spammers. Zhu et al. [29] found that spammers are separated in Renren, which is a Facebook-like social network. Yang et al. [26] had an opposite finding in Twitter, where spammers tend to be inter-connected, possibly trying to camouflage each other. Consequently, different algorithms for spammer detection were proposed for the two networks [14, 29].

It is commonly agreed that legitimate users favor only other legitimate users, and they do not follow others at random. For example, Weng et al. [22] found that the presence of reciprocity in Twitter can be explained by the theory of "homophily". Users sharing similar topics are more likely to follow each other reciprocally. Hopcroft et al. [12] showed strong evidence of the structural balance among reciprocal relationships, i.e. users with common friends of reciprocal ties have a tendency to follow each other. The above findings indicate that some users do follow others "seriously".

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
*WWW 2015 Companion,* May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2745400.


**Figure 1: Example of "careless" users when following others**

On the other hand, evidence of legitimate users following spammers was also found. Ghosh et al. [8] 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. [26] also observed similar users, which in turn aid spammers to spread spams and avoid detection. More than the above, we found a significant number of legitimate users following more spammers than expected.

The above 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, e.g. social capitalists. 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 types of user. 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 towards 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 of spammer detection [5, 6, 24] assume that legitimate users favor only other legitimate users. We avoid such assumptions by leveraging the proposed carefulness. For example, as shown in Figure 1, the users themselves are legitimate, but they appear to follow back anyone who follows them, so they are potentially following spammers.

Given the carefulness of users, the second question is that: How can it be leveraged to aid spammer detection?

In this paper, 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 features proposed in previous works for spammer detection, and illustrate how the carefulness is incorporated to improve the detection.
- We evaluate our method on a real dataset with millions of users. Our results show that our method is able to characterize user behavior in terms of the carefulness and it is effective to detect spammers.
- We illustrate how other applications (e.g. link prediction) can benefit from the proposed carefulness.

In the rest of this paper, 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). Evaluation of our approach is presented in Section 6. Several technical issues and other applications are discussed in Section 7. Finally, we conclude our results and discuss future works based on the proposed method (Section 8).

## 2. RELATED WORK

Spammer detection in social networks, e.g. email systems [4, 6] and SMS networks [23], has been widely studied for many years. In recent years, spammers in microblogging websites have attracted increasing attentions from researchers and developers. Previous works mainly focus on characterizing abnormal or spamming behaviors in various aspects [8, 9, 20, 26, 27]. Another major topic is detecting spammers based on content of tweets, network structure, or both. In this paper, we focus on detecting spammers based on network structure.

Benevenuto et al. [3] studied the problem of spammer detection in Twitter. They analyzed the tweet content and user social activities in Twitter, from which they extracted a number of features for detection. Hu et al. [13, 14] 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. [29] also employed a matrix factorization approach for Facebook-like social networks. They made a different assumption about neighboring users: While legitimate users are inter-connected, 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.

The above approaches require additional information other than the network structure, and seem to underestimate the knowledge of legitimate users. As shown in this paper, 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.

A number of works adapt PageRank and its variants to rank spammers based on the graph structure. Gyöngyi et al. [10] proposed TrustRank to detect web spams. TrustRank is initiated with a set of known good websites as seeds, and then propagates the scores with biases. Chirita et al. [6] proposed to rank the reputation of email senders with a variant of PageRank. Cao et al. [5] employed an idea of early-terminated random walks to detect fake accounts in online social networks. Xue et al. [24] further utilized the information of friend requests to enhance the detection. In a microblogging website, a link is usually established without the permission of users, so their approach appears not applicable in this case.

In general, PageRank based methods assume that legitimate users favor only other legitimate users. However, the case that some legitimate users follow spammers [8, 26], which occurs quite often, is not considered. To address this problem, we propose the carefulness to characterize such behavior separately. It is one of the main differences between our method and previous works.

## 3. PRELIMINARY

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

**Definition** We model users and their social ties in a microblogging website as a directed social graph $G = (V, E)$. Every node in $V$ corresponds to a unique user in the website. A directed edge $(u, v) \in E$ is presented in the graph if and only if the user $u$ is following $v$. The edges $(u, v)$ 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 a user $u$ are represented as a set $N_O(u) = \{v|(u, v) \in E\}$. Additionally, users who have reciprocal relations with $v$ are denoted as $N_R(v) = N_I(v) \cap N_O(v)$. In the rest of this paper, we would use the terms "user" and "node" interchangeably.

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

## 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

So 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 websites in China, to seek answers for this question.

Our dataset contains 3.5 million users and 2,000 users are manually identified as legitimate user or spammer (see Section 6.3). 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. In our dataset, 8.4% 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% legitimate users and 68.9% spammers follow at least one suspended user. More importantly, 11.5% legitimate users and 23.7% 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 it is quite often for legitimate users to follow spammers. 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 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.

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

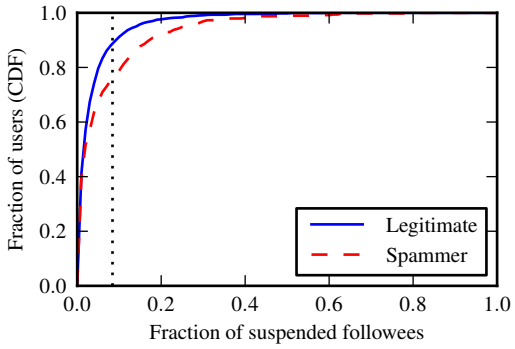



**Figure 2: Cumulative distributions of the fraction of suspended followees for legitimate users and spammers. The vertical line denotes the expected fraction of suspended followees if a user follows others at random.**

### 4.2 Carefulness

We define the carefulness as the probability of $u$ identifying a legitimate user or a spammer correctly. To simplify the problem, we assume 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 above observation suggests a possible approach that it can be inferred from one's followees. With a handful of spammers identified by experts, we build the connection between the carefulness of a user and the labels of her followees. 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 $u$ decides whether to follow $v$ based on her knowledge of $v$. User $u$ is assumed to follow only users that she considers as legitimate. However, considering $v$ as legitimate is not enough for an actual "follow" action. For example, it is also determined by various properties of the network and users, e.g. proximity [18], homophily [22], and structural balance [12]. Given that $u$ considers $v$ legitimate, we define $r(u, v)$ as the 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

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

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 towards legitimate users or spammers ($f(u) = 1/2$)
- *Malicious* users who always seek for spammers and pay no attention to legitimate users ($f(u) = 0$)

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

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

Given an existing edge $(u,v)$, Equation (2) shows that $f(u)$ only depends on whether the followees are legitimate, which means 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 less 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 $X_u = (x_{u1}, x_{u2}, ..., x_{uk})$ associated with $u$:

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

The features are described in Section 5. A dummy feature $x_{u0} = 1$ is included to make $w_0$ an intercept. In this paper, we only focus on structural features (e.g. degrees), and leave the use of user profiles and tweets for future works.

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 the 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 $v$ is a spammer. As malicious users are seeking for spammers on purpose, it is unlikely for a legitimate user to gain so much attention from them.

Regarding the above requirements, we find the average of $P(Y_v = 1 | (u,v) \in E)$ as a good choice:

$$\begin{aligned} g(v) &= \frac{1}{|N_I(v)|} \sum_{u \in N_I(v)} P(Y_v = 1 | (u,v) \in E) \\ &= 1 - \frac{1}{|N_I(v)|} \sum_{u \in N_I(v)} d(u) \end{aligned} \quad (4)$$

The prior probability $P(Y_v = 1)$ 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, so $d(u)$ is defined as

$$d(u) = \frac{1}{1 + \frac{p_s}{1 - p_s} \exp\left(-\sum_{i=0}^{k} w_i x_{ui}\right)} \quad (5)$$

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:

$$L(\mathbf{w}) = \frac{1}{2} \sum_{v \in D} (g(v) - y_v)^2 + \frac{\lambda}{2} \sum_{i=0}^{k} w_i^2 \quad (6)$$

We discuss other choices of the loss function in Section 7.1.

## 4.4 Training the Model

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

$$\frac{\partial L(\mathbf{w})}{\partial w_i} = \sum_{v \in D} (g(v) - y_v) \frac{\partial g(v)}{\partial w_i} + \lambda w_i \quad (7)$$

Taking the derivative of $g(v)$ gives

$$\begin{aligned} \frac{\partial g(v)}{\partial w_i} &= -\frac{1}{|N_I(v)|} \sum_{u \in N_I(v)} \frac{\partial d(u)}{\partial w_i} \\ &= -\frac{1}{|N_I(v)|} \sum_{u \in N_I(v)} d(u)(1 - d(u)) \cdot x_{ui} \end{aligned} \quad (8)$$

We apply a gradient descent algorithm to solve the optimization problem (Algorithm 1). All features are standardized for better convergence. The gradient descent algorithm may probably stuck in a local minimum, so we repeat the algorithm several times with different starting points to find a good solution. 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.

**Algorithm 1** Learning the carefulness

**Input:** Social graph $G = (V, E)$, known spammers and legitimate users $D$, parameter $\lambda$, learning rate $\eta$
**Output:** Learned paramter $\mathbf{w}$

$\mathbf{w}^{(0)} \leftarrow$ random values
$t \leftarrow 0$
**repeat**
  **for each** $u \in V$ **do**
    Calculate $d(u)$ with $\mathbf{w}^{(t)}$ ▷ Equation (5)
  **end for**
  **for each** $v \in D$ **do**
    Calculate $g(v)$ with $\mathbf{w}^{(t)}$ ▷ Equation (4)
    Calculate $\frac{\partial g(v)}{\partial \mathbf{w}}$ with $\mathbf{w}^{(t)}$ ▷ Equation (8)
  **end for**
  $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \frac{\partial L(v)}{\partial \mathbf{w}}$ ▷ Equation (7)
  $t \leftarrow t + 1$
**until** convergence
**return** $\mathbf{w}^{(t)}$

# 5. DETECTING SPAMMERS

Given the carefulness $f(u)$, it is still unclear how it can be leveraged to detect spammers in a microblogging website. 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, e.g. reciprocity and communities. It also becomes unreliable as the number of followers decreases. Herein, we propose an approach that incorporates structural features capturing user behaviors in various aspects. We first review a set of features that were proposed in previous works for spammer detection. We then describe how they can be adjusted based on the carefulness. We refer to the two versions of features as the *original* and the *adjusted* respectively. Note that the original version of features is used in learning the carefulness.

## 5.1 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 would follow back. In [15], Huang et al. 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 website, we define the *follow-back rate* as $|N_R(v)|/|N_O(v)|$ analogously. As a user tends to follow legitimate users, the follow-back rate of a spammer is expected to be low.

**Adjustment** However, these features can be easily manipulated by creating fake accounts and reciprocal relations between them. We propose the adjusted follow-back rate as $\sum_{u\in N_R(v)} f(u)/|N_O(v)|$ to avoid fake links. A malicious user who follows back can not help to manipulate this feature. A legitimate user gets a much higher follow-back rate because she is favored by careful users. We apply the similar strategy to adjust other degree features as the sums of $f(u)$.

## 5.2 Clustering Coefficients

In [4], Boykin and Roychowdhury 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 $V'$:

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

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

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

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

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), while $C_R(u)$ is limited to communities that $u$ actually belongs to. Given a legitimate user $u$, $C_O(u)$ tends to be less than $C_R(u)$, since the user may be following several irrelevant communities. 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 numbers of reciprocal relations in a neighborhood, we adjust them by counting only "real" links:

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

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

The above adjustment makes the clustering coefficients of spammers even lower than those of legitimate users. Particularly, if a spammer manages to make a few dense fake communities, the adjusted clustering coefficients are still low, because the carefulness of members are expected to be low.

## 5.3 PageRank

PageRank and its variants are widely used in ranking web pages. In recent works [5, 6, 15, 24], 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 out-going nodes with a damping factor $d$:

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

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 out-going 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.

**Adjustment** One drawback of PageRank based methods is that a spammer can still get a high score if she manages to attract a few legitimate users. We fix this by introducing personalized damping factors. For a careful user, we shall walk towards her followees with a high probability, since she knowns them legitimate with a high confidence. On the other hand, 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 $f(u)$:

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

When a node $u$ is arrived at, the random walk follows edges 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.

## 5.4 Classification

We model the detection of spammers 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.

# 6. EXPERIMENTS

In this section, we start by introducing the dataset and ground-truth for evaluation. 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 $f(u)$. Finally, we discuss the selection of parameters and efficiency issues.

## 6.1 Dataset

We used Sina Weibo as the data source for our evaluation. We crawled our dataset in May, 2014 using the API of Sina Weibo. We applied the following strategy to obtain a reasonably "good" sample [17] from the whole website. We first sampled a number of tweets posted during April and May, 2014, from the public timeline of the website, 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. In other words, we crawled the 2-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 edges, among which 83 million pairs of users follow each other reciprocally. The social graph is connected except for a few dozens of isolated nodes.

In previous works, various criteria were used to identify spammers for ground-truth, e.g. suspended accounts [14], unrelated tweets and hashtags [3], social honeypots [16, 19], and malicious URLs [26]. It should be noted that these criteria may be biased to certain types of spammer. In this paper, we are 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. Suspended users 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, e.g. only a few tweets without actual content were posted.

During the inspection, we spotted (but not limited to) several typical patterns of spammers. A significant number of spammers post snippets from online news or blogposts, possibly trying to avoid content based detection. We consider such users as spammers because they occasionally post URLs to malicious websites or irrelevant online shops. This behavior is quite different from (legitimate) regular marketers whose tweets are mostly relevant to their products. Some other spammers go further by copying personal tweets (e.g. "my cat is sick") and photos from other users, making them more similar to real users. Such activities are identified by searching for those tweets and photos in the website, and comparing watermarks in photos and timestamps. 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 a number of fake accounts and 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. [28] 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 the above patterns do not cover all spammer. A significant number of spammers do not exhibit such obvious patterns and require human comprehension to identify. 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. [28], a large fraction of trends in Sina Weibo are actually artificially inflated by fake accounts. We used a 10-fold cross-validation in all experiments and reported the averages. In each fold, 90% of the labels were used to learn the carefulness $f(u)$ and to train the classifier. The remaining was for testing.

## 6.2 Carefulness

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

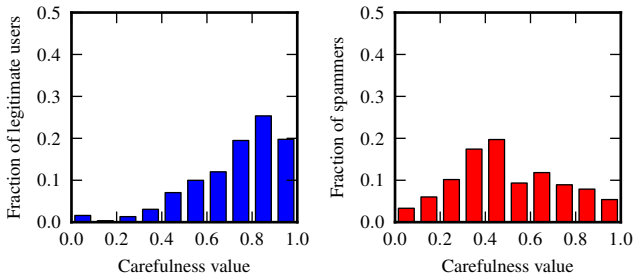
### 6.2.1 Spammers

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. In general, the result (Figure 3) shows a 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% legitimate users. We also find that most legitimate users concentrate in range $[0.8, 0.9]$ but relatively few of them are extremely careful, e.g. $f(u) > 0.9$ for 20% legitimate users. This is consistent with our observation in Section 4.1 that a large fraction of legitimate users follow at least one spammer.

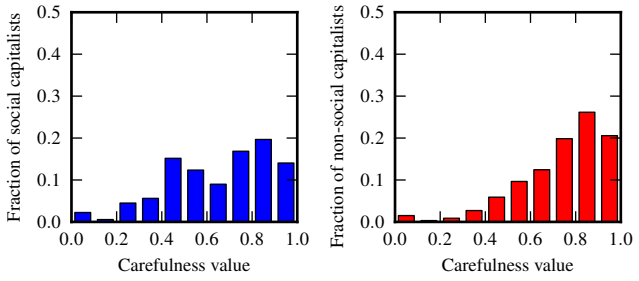
On the other hand, spammers show various carefulness in all range. Most spammers appear to be careless, and the others are either malicious or careful. This could be explained by the different strategies of spammers to seek user IDs. Most spamming accounts are controlled by automated scripts, so they follow whoever they see, making them appear to be careless. Some accounts of spammers are used to boost the reputation of other spammers [26] or paid users [28], so they behave either maliciously or carefully.

While the carefulness is learned from users' followees rather than themselves, the above results show the correlation between it and the type of users. The result is roughly consis-




(a) Legitimate users (b) Spammers

**Figure 3: Distributions of the carefulness for legitimate users and spammers**




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

**Figure 4: Distributions of the carefulness for social capitalists and non-social capitalists**

tent 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.

### 6.2.2 *Social capitalists*

As shown by Ghosh et al. [8], 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. We identified social capitalists from known legitimate users as follows. We obtained the category of a user, e.g. civilians, famous artists, or enterprises, via the API of Sina Weibo[2]. We considered a user as a social capitalist, if she was not 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, while non-social capitalists simply subscribe to popular accounts and communicate with friends. We ended up with 12.4% legitimate users as social capitalists.

The distribution of social capitalists (Figure 4(a)) shows two distinct peaks. 28% 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 other peak, 34% social capitalists have a carefulness value greater than 0.8. We inspected social capitalists with top carefulness value 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, joke, and life style. 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 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 Figure 4(b) is expected to be similar to Figure 3(a), because the majority of legitimate users are non-social capitalists.

[2]https://api.weibo.com/2/users/show.json

**Table 1: Average carefulness of users grouped by profile features**

| Feature | Average carefulness | |
|---|---|---|
| | Yes | No |
| Posted any tweet? | **0.700** | 0.682 |
| Tweet in favorite? | **0.750** | 0.603 |
| Non-empty bio? | **0.728** | 0.638 |
| Custom domain? | **0.735** | 0.699 |
| Personal website? | **0.717** | 0.696 |
| Direct message from stranger? | 0.663 | **0.704** |
| Comment from stranger? | 0.696 | **0.733** |
| Is geo-location enabled? | 0.697 | **0.740** |
| Is verified? | **0.700** | 0.549 |

### 6.2.3 *Profiles*

We also crawled the profiles of users in our dataset. We extracted three groups of binary features from the profiles, focusing on inactive users, privacy setting, 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 1) shows that active users, i.e. those who have ever posted a tweet, saved a tweet in favorite, written a bio, applied for a custom domain, or specified the URL to personal website, are more careful. Active users learn about spammer's strategies while browsing the website, so they are better at avoiding spammers. We also find that users who are more concerned about privacy, i.e. disallowing direct messages or comments from strangers, or hiding locations, are inferred as more careful. This is reasonable because these users are not likely to follow others at random, or 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 website, so they ought to maintain their accounts seriously. In summary, although the carefulness is learned based on graph structure, these results show its interesting correlation with profiles.

## 6.3 Detection

Now we detect spammers based on the adjusted features. We start with single features, and then we combine them together for the best performance. We also compare our results with the current detection system of Sina Weibo. The selection of parameter is discussed in the end.

### 6.3.1 *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. The true positive rate (TPR) is

**Table 2: Accuracy of detection with individual features**

| Feature | AUC | | Gain |
|---|---|---|---|
| | Original | Adjusted | |
| Number of followees | 0.578 | **0.673** | 16.4% |
| Number of followers | 0.612 | **0.715** | 16.8% |
| Number of reciprocal relations | **0.734** | 0.673 | −8.3% |
| Follow-back rate | 0.714 | **0.786** | 10.1% |
| Clustering coefficient $C_O(u)$ | 0.848 | **0.858** | 1.2% |
| Clustering coefficient $C_R(u)$ | 0.780 | **0.820** | 5.1% |
| PageRank | 0.666 | **0.756** | 13.5% |

defined as the fraction of correctly identified spammers out of actual spammers. The false positive rate (FPR) is defined as the fraction of legitimate users that are misclassified out of actual legitimate users. The trade off between TPR and FPR can be visualized by the receiver operating characteristic (ROC). We quantify the overall performance with the area under the curve (AUC).

### 6.3.2 Features

We used each feature described in Section 5 alone to detect spammers, and compared the performance of the original version and the adjusted version. The result (Table 2) shows consistent improvements over the original ones, except for the number of reciprocal relations.

Degree features can be easily manipulated by spammers by connecting fake accounts. These features are expected to work poorly at the first place. When they are adjusted with the carefulness, a significant improvement occurs. The adjusted number of reciprocal relations is shown to be less effective, but the performance drop is relatively small compared to other degree features. The follow-back rate is also improved by counting only seriously established follow-backs. Clustering coefficients are the most effective for detection by themselves. A slight improvement is observed by avoiding fake communities. The performance boost over PageRank is quite surprising. We believe it is mainly due to replacing the global damping factor with personalized ones.

### 6.3.3 Evaluation

We combined all adjusted features and used Random Forests to perform the detection (`RF-adjusted`). In our initial experiments, we tried several other 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, so we chose it in our experiments. For comparisons, we employed the following methods as baselines:

- We use the original version of features to train another Random Forests (`RF-original`) to examine the effect of the carefulness.
- 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 [10] 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 inverse PageRank (i.e. inversing the direction of edges), and uniform sampling. It turned out uniform sampling yields the best performance, so we took a sample of 100 legitimate users as seeds.

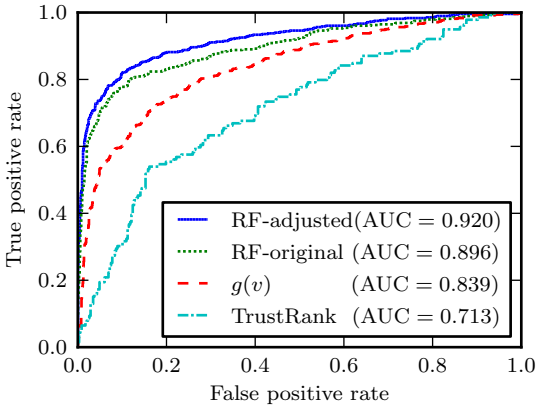



**Figure 5: ROC curves of different detection methods**

**Table 3: Accuracy of detection with stacked features**

| Features | AUC | Gain |
|---|---|---|
| Degrees | 0.902 | N/A |
| +Clustering coefficients | 0.918 | +1.8% |
| +PageRank | 0.920 | +0.2% |

In addition to the above baselines, we also considered matrix factorization based methods recently proposed by Zhu et al. [29] and Hu et al. [14]. 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 above two methods cannot work properly, so we do not compare with them here.

The result (Figure 5) shows that the estimated label $g(v)$ outperforms TrustRank significantly. This confirms our observation that a legitimate user is not always following legitimate users, and it is necessary to model the carefulness separately. Random Forests with a rich set of features outperform the single function $g(v)$. This 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, while TrustRank and $g(v)$ are unable to capture such patterns. By adjusting features with the proposed carefulness, the performance is further improved. Particularly, at a low FPR of 1%, the TPR is reasonably high (53%). This is a promising result, since misclassifying legitimate users as spammers is usually expensive and should be avoided in a real detection system.

As the original version of features treats every edge equally, it could be manipulated by fake social relations. By weighting edges with the carefulness, such an effect is reduced by a considerable extent, making the features more effective for the detection. We further evaluated the performance by adding features group by group. The result (Table 3) shows consistent improvement with the addition of information, indicating that all features contribute to the detection.

We conducted an online test to verify our results. We sampled two groups of identified spammers 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 spammers were suspended, while 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. While we may keep reporting the first group, the result indicates that those spammers were difficult for the website to detect currently, and our method is effective in capturing such spammers.

#### 6.3.4 Parameter

The parameter $\lambda$ (see Equation (6)) trades off between the complexity and the fitness of the model. We evaluated the overall performance of our method with various values of $\lambda$ $(10^{-6}, 10^{-5}, ..., 10^{6})$. The result (Figure 6) shows that the performance is not sensitive for $\lambda \leq 10^3$ and drops significantly when $\lambda$ goes above $10^3$. A smaller value of $\lambda$ would increase the effect of over-fitting. On the other hand, as $\lambda$ increases, the loss function $L(\mathbf{w})$ is dominated by the regularization term $\frac{\lambda}{2}\sum_{i=0}^{k} w_i^2$, resulting in $\mathbf{w} \approx 0$ and $f(u) \approx 1/2$. So we choose $\lambda = 1$ as a robust choice.

#### 6.3.5 Efficiency

We implemented our algorithm in C++. The experiments were performed on a server equipped with an Intel Xeon E5-2665 CPU. While holding the full graph takes 5 GB memory, our algorithm itself actually did not need much memory. We ran the algorithm with 32 different random start points in parallel, and we picked the one with minimum loss $L(\mathbf{w})$ as the final solution. The learning process was terminated when the improvement of $L(\mathbf{w})$ is less than 0.1. On average, it took about 40 iterations and 10 minutes to converge. Extracting the features needed another 10 minutes. In total, performing a detection on our dataset took 20 minutes, which was reasonably efficient.

## 7. 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.

### 7.1 Loss Function

The choice of loss function plays an important role in most machine learning algorithms. As the learned carefulness is incorporated with various features, 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 (MLE), i.e. seeking the value of $\mathbf{w}$ 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 features.

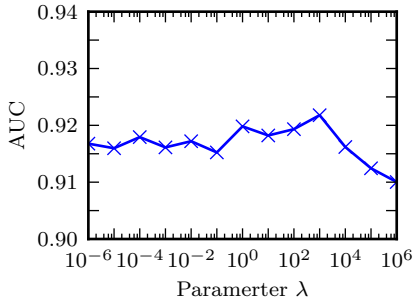



**Figure 6: Accuracy of detection with various values of $\lambda$**

Another approach is adapted from the work of Backstrom and Leskovec [2]. 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. While this requirement is too hard to satisfy in practice, it is relaxed with a loss function as

$$L(\mathbf{w}) = \sum_{u \in D^+, v \in D^-} h(g(v) - g(u)) + \frac{\lambda}{2}\sum_{i=1}^{k} w_i^2 \qquad (9)$$

where $h(x) = (1 + \exp\{-x/b\})^{-1}$ is the Wilcoxon-Mann-Whitney (WMW) loss [25] 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 be a problem when the training set becomes larger. Therefore, we use the squared error loss as the best choice in our method.

### 7.2 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 [29], and 0-day spammers that are never observed before [16]. 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 $r(u)$. We have the probability of creating an edge $(u, v)$ as

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

We assume that the observed social graph $G$ is generated based on this 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, which is better than any single unsupervised feature in Table 2.

### 7.3 Link Prediction

In a microblogging social network, the link prediction problem [18] 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, thus it is more likely for them to form a connection. However, if the common friends are careless or even malicious, we would expect 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 the following typical methods to predict if $u$ will follow $v$, and introduce how to adjust them based on the above 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 $S$.

**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, i.e. $f(N_O(u) \cap N_I(v))$.

**Jaccard's coefficient** The Jaccard's 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 [1] 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 [2, 18] 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 re-define the restart probability with the carefulness similarly to Section 5.3. When the random walk arrives at a node $w$, it returns to $u$ with probability $1 - f(w)$.

We focus on predicting links between nodes that are 2-hops from a given node [2]. A node pair $(u, v)$ is considered as a positive sample if there is an edge 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 that are both known legitimate.

We find that the 2,000 labeled users mentioned in Section 6.1 are mostly apart from each other, resulting in insufficient number of testing samples. So 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. We measure the performance of prediction by the area under the curve (AUC).

The result (Table 4) shows consistent improvements over the original methods. The adjusted random walk yields the best performance overall. 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 the carefulness is only helpful for real users.

**Table 4: Accuracy of link prediction**

| Measure | AUC | | |
|---|---|---|---|
| | Original | Adjusted | Gain |
| Common friends | 0.761 | **0.782** | 2.8% |
| Jaccard's coefficient | 0.678 | **0.688** | 1.5% |
| Adamic/Adar | 0.786 | **0.787** | 0.1% |
| Preferential attachment | 0.563 | **0.571** | 1.4% |
| Random walk | 0.948 | **0.964** | 1.7% |

Various additional information, e.g. graph attributes [2], contents [7], and locations [21], has been shown 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 to 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.

# 8. CONCLUSION

As the behavior of users varies when they are following someone else in a microblogging website, 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 then illustrate how spammer detection can be enhanced using the proposed measure. Our experiments show that the carefulness is indeed effective for the detection.

There are many potential future works based on this paper. It would be interesting to combine the content information, e.g. tweets, photos, and profiles, to enhance the inference of carefulness. The carefulness itself can be used as tool to analyze and interpret user behaviors in a microblogging website. It would also be interesting to apply the proposed method to other types of social networks, e.g. email communication networks.

The proposed model of carefulness can be extended to capture more fine-grained patterns. Similarly to most spammer detection systems, the false positive rate and false negative rate of a user are not necessarily the same. While 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 pair-wise 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 works.

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, since 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 think characterizing users themselves and leveraging their power to detect spams is a promising direction towards this problem.

# 9. REFERENCES


[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] L. Backstrom and J. Leskovec. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 635–644, 2011.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on Twitter. In *Collaboration, Electronic messaging, Anti-abuse and Spam conference*, volume 6, page 12, 2010.

[4] P. Boykin and V. Roychowdhury. Leveraging social networks to fight spam. *Computer*, 38(4):61–68, 2005.

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

[6] P.-A. Chirita, J. Diederich, and W. Nejdl. Mailrank: Using ranking for spam detection. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 373–380, 2005.

[7] S. Gao, L. Denoyer, and P. Gallinari. Temporal link prediction by integrating content and structure information. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1169–1174, 2011.

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

[9] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, pages 27–37, 2010.

[10] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proceedings of the 30th International Conference on Very Large Data Bases*, pages 576–587, 2004.

[11] P. Heymann, G. Koutrika, and H. Garcia-Molina. Fighting spam on social web sites: A survey of approaches and future challenges. *Internet Computing, IEEE*, 11(6):36–45, 2007.

[12] J. Hopcroft, T. Lou, and J. Tang. Who will follow you back?: Reciprocal relationship prediction. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1137–1146, 2011.

[13] X. Hu, J. Tang, and H. Liu. 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*, pages 547–556, 2014.

[14] X. Hu, J. Tang, Y. Zhang, and H. Liu. Social spammer detection in microblogging. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2633–2639, 2013.

[15] J. Huang, Y. Xie, F. Yu, Q. Ke, M. Abadi, E. Gillum, and Z. M. Mao. 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*, pages 143–148, 2013.

[16] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 435–442, 2010.

[17] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining*, pages 631–636, 2006.

[18] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proceedings of the 12th International Conference on Information and Knowledge Management*, pages 556–559, 2003.

[19] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 1–9, 2010.

[20] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: An analysis of Twitter spam. In *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, pages 243–258, 2011.

[21] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining*, pages 1100–1108, 2011.

[22] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: Finding topic-sensitive influential twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 261–270, 2010.

[23] Q. Xu, E. Xiang, Q. Yang, J. Du, and J. Zhong. SMS spam detection using noncontent features. *IEEE Intelligent Systems*, 27(6):44–51, Nov 2012.

[24] J. Xue, Z. Yang, X. Yang, X. Wang, L. Chen, and Y. Dai. Votetrust: Leveraging friend invitation graph to defend against social network sybils. In *Proceedings of the 32nd IEEE International Conference on Computer Communications*, pages 2400–2408, 2013.

[25] L. Yan, R. H. Dodier, M. Mozer, and R. H. Wolniewicz. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In *Proceedings of the 20th International Conference on Machine Learning*, pages 848–855, 2003.

[26] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu. 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*, pages 71–80, 2012.

[27] S. Yardi, D. Romero, and G. Schoenebeck. Detecting spam in a Twitter network. *First Monday*, 15(1), 2009.

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

[29] Y. Zhu, X. Wang, E. Zhong, N. N. Liu, H. Li, and Q. Yang. Discovering spammers in social networks. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012.