Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Social Spammer Detection in Microblogging

Xia Hu[†], Jiliang Tang[†], Yanchao Zhang[‡], Huan Liu[†]
[†]Computer Science and Engineering, Arizona State University, USA
[‡]School of Electrical, Computer, and Energy Engineering, Arizona State University, USA
{xiahu, jiliang.tang, yczhang, huan.liu}@asu.edu

Abstract

The availability of microblogging, like Twitter and Sina Weibo, makes it a popular platform for spammers to unfairly overpower normal users with unwanted content via social networks, known as social spamming. The rise of social spamming can significantly hinder the use of microblogging systems for effective information dissemination and sharing. Distinct features of microblogging systems present new challenges for social spammer detection. First, unlike traditional social networks, microblogging allows to establish some connections between two parties without mutual consent, which makes it easier for spammers to imitate normal users by quickly accumulating a large number of "human" friends. Second, microblogging messages are short, noisy, and unstructured. Traditional social spammer detection methods are not directly applicable to microblogging. In this paper, we investigate how to collectively use network and content information to perform effective social spammer detection in microblogging. In particular, we present an optimization formulation that models the social network and content information in a unified framework. Experiments on a real-world Twitter dataset demonstrate that our proposed method can effectively utilize both kinds of information for social spammer detection.

1 Introduction

Microblogging, like Twitter and Sina Weibo, has become a widely popular platform for information dissemination and sharing in various scenarios such as marketing, journalism or public relations. With the growing availability of microblogging, social spamming has become rampant. Many fake accounts, known as social spammers [Webb *et al.*, 2008], are employed to unfairly overpower normal users. Social spammers can be coordinated to launch various attacks such as befriending victims and then grabbing their personal information [Bilge *et al.*, 2009], conducting spam campaigns which lead to phishing, malware, and scams [Grier *et al.*, 2010], and conducting political astroturf [Ratkiewicz *et al.*, 2011a; 2011b]. Successful social spammer detection in microblogging presents its significance to improve the quality of user experience, and to positively impact the overall value of the social systems going forward [Lee *et al.*, 2010].

Spammer detection has been studied in various online social networking (OSN) platforms, e.g., Youtube [O'Callaghan *et al.*, 2012] and Facebook [Brown *et al.*, 2008]. One effective way to perform spammer detection is to utilize the social network information [Boykin and Roychowdhury, 2005; Danezis and Mittal, 2009]. This scheme is built upon the assumption that spammers cannot establish an arbitrarily large number of social trust relations with normal users. However, different from other OSNs, microblogging systems feature unidirectional user binding, meaning anyone can follow anyone else without prior consent from the followee.[1] Many users simply follow back when they are followed by someone for the sake of courtesy [Weng *et al.*, 2010]. Due to the *reflexive reciprocity*, it is easier for spammers to imitate normal users in microblogging by quickly accumulating a large number of social relations. A recent study [Ghosh *et al.*, 2012] on microblogging shows that spammers can successfully acquire a number of normal followers, especially those referred to as social capitalists who tend to increase their social capital by following back anyone who follows them.

Meanwhile, microblogging provides additional content information, i.e., microblogging messages, other than the social networks. Different from email spam detection, content analysis in microblogging for social spammer detection has been little studied due to the distinct characteristics of microblogging messages. First, microblogging messages are very short. For example, Twitter allows users to post messages up to 140 characters. Short messages bring new challenges to traditional text analytics. They cannot provide sufficient context information for effective similarity measure, the basis of many text processing methods [Hu *et al.*, 2009]. Second, microblogging messages are very unstructured and noisy. In particular, when composing a message, users often prefer to use newly created abbreviations or acronyms that seldom appear in conventional text documents. For example, messages like "How r u?" and "Good 9t" are popular in

[1]Although there is often an option for a user to manually (dis)approve a following request, it is rarely used by normal users for convenience.

microblogging systems, but they are not even formal words. Although they provide a better user experience, unstructured expressions make it very difficult to accurately identify the semantic meanings of these messages. Last, microblogging messages are *networked* [Zhu *et al.*, 2012] in the sense that they are generated by users following some others in microblogging systems. The traditional assumption in many applications that data instances are independent and identically distributed (i.i.d.) is thus no longer valid for networked microblogging messages. The distinct features make traditional text analytics less applicable in microblogging platforms.

To address the new challenges posed by microblogging services, we propose to take advantage of both network and content information for social spammer detection in microblogging. In this paper, we study the problem of social spammer detection in microblogging with network and content information. In essence, we investigate: (1) how to model the network information and content information properly in microblogging; and (2) how to seamlessly utilize both sources of information for the problem we are studying. Our solutions to these two challenges result in a new framework for Social Spammer Detection in Microblogging (*SSDM*). In particular, we employ a directed Laplacian formulation to model the refined social networks, and then integrate the network information into a sparse supervised formulation for the modeling of content information. The main contributions of the paper are outlined as follows:

- We formally define the problem of social spammer detection in microblogging with both network and content information;
- We propose a unified model to effectively integrate both social network information and content information for the problem we are studying; and
- We empirically evaluate the proposed framework on a real-world Twitter dataset and elaborate the effects of each type of information for social spammer detection.

The remainder of this paper is organized as follows. In Section 2, we formally define the problem of social spammer detection in microblogging. In Section 3, we propose a supervised model to integrate both network and content information for spammer detection. In Section 4, we report empirical results on a real-world dataset. In Section 5, we conclude and present the future work.

2 Problem Formulation

In this section, we first introduce the notations used in the paper and then formally define the problem we study.

Notation: The following notations are used. Matrices are denoted by boldface uppercase letters, vectors by boldface lowercase letters, and scalars by lower case letters. Let $\|\mathbf{A}\|$ denote the Euclidean norm, and $\|\mathbf{A}\|_F$ the Frobenius norm of the matrix $\mathbf{A}$. Specifically, $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{A}_{ij}^2}$. Let $\mathbf{A}^T$ denote the transpose of $\mathbf{A}$.

Let $\mathbf{U} = [\mathcal{G}, \mathbf{X}, \mathbf{Y}]$ be a target microblogging user set with social network information $\mathcal{G}$, content information of microblogging messages $\mathbf{X}$, and identity label matrix $\mathbf{Y}$. We use $\mathcal{G} = (V, E)$ to denote the social network, where nodes u and v in V represent microblogging users, and each directed edge $[u, v]$ in E represents a following relation from u to v. We do not have self links in the graph, i.e., $u \neq v$. We use $\mathbf{X} \in \mathbb{R}^{m\times n}$ to denote content information, i.e., messages posted by the users, where m is the number of textual features, and n is the number of users. We use $\mathbf{Y} \in \mathbb{R}^{n\times c}$ to denote the identity label matrix, where c is the number of identity labels. Following previous work on spammer detection [Benevenuto *et al.*, 2010; Lee *et al.*, 2010], we focus on classifying users as either spammers or normal users, i.e., $c = 2$. It is straightforward to extend this setting to a multi-class classification task.

With the given notations, we formally define the problem of social spammer detection in microblogging as follows:

Given a set of microblogging users $\mathbf{U}$ *with social network information* $\mathcal{G}$*, content information* $\mathbf{X}$*, and identity label information* $\mathbf{Y}$ *of part of the users in the set (i.e., training data), we aim to learn a classifier* $\mathbf{W}$ *to automatically assign identity labels for unknown users (i.e., test data) as spammers or normal users.*

3 Social Spammer Detection

In this section, we first introduce how we model social network information for spammer detection and then discuss the modeling of microblogging messages for each user. Finally, we present a framework *SSDM* that considers both network and content information with its optimization algorithm.

3.1 Modeling Social Network Information

To make use of network information, many methods assume that two nodes share a similar label when they are mutually connected in the network [Chung, 1997; Gu and Han, 2011; Zhu *et al.*, 2012]. It has distinct features in microblogging. First, users have a directed following relation in microblogging. Second, spammers can easily follow a large number of normal microblogging users within a short time. Thus the existing methods are not suitable to this problem.

We first refine the social relations in the social network. Given the social network information $\mathcal{G}$ and the identity label matrix $\mathbf{Y}$, we have four kinds of following relations: [spammer, spammer], [normal, normal], [normal, spammer], and [spammer, normal]. Since the fourth relation can be easily faked by spammers, we make use of the first three relations in the proposed framework. Now we introduce how to represent and model the social network information in detail.

The adjacency matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$ is used to represent the refined social network $\mathcal{G}$, and it is defined as

$$\mathbf{G}(u,v) = \begin{cases} 1 & \text{if [u, v] is among the first three relations} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where u and v are nodes, and $[u, v]$ is a directed edge in the graph $\mathcal{G}$. The in-degree of the node u is defined as $\mathbf{d}_u^{in} = \sum_{[v,u]}\mathbf{G}(v,u)$, and the out-degree of the node u is defined as $\mathbf{d}_u^{out} = \sum_{[u,v]}\mathbf{G}(u,v)$. Let $\mathbf{P}$ be the transition probability matrix of random walk in a given graph with $\mathbf{P}(u,v) = \mathbf{G}(u,v)/\mathbf{d}_u^{out}$ [Zhou *et al.*, 2005]. The

random walk has a stationary distribution $\boldsymbol{\pi}$, which satisfy $\sum_{u \in V} \boldsymbol{\pi}(u) = 1$, $\boldsymbol{\pi}(v) = \sum_{[u,v]} \boldsymbol{\pi}(u)\mathbf{P}(u,v)$ [Chung, 2005; Zhou *et al.*, 2005], and $\boldsymbol{\pi}(u) > 0$ for all $u \in V$.

The key idea here is that we employ network information to smooth the learned model. It can be mathematically formulated as minimizing

$$\mathcal{R}_S = \frac{1}{2} \sum_{[u,v] \in E} \boldsymbol{\pi}(u)\mathbf{P}(u,v) \|\hat{\mathbf{Y}}_u - \hat{\mathbf{Y}}_v\|^2, \quad (2)$$

where $\hat{\mathbf{Y}}_u$ denotes the predicted label of user u, and $\hat{\mathbf{Y}}_v$ the predicted label of user v. The loss function will incur a penalty if two users have different predicted labels when they are close to each other in the graph.

Let $\boldsymbol{\Pi}$ denote a diagonal matrix with $\boldsymbol{\Pi}(u,u) = \boldsymbol{\pi}(u)$.

Theorem 1 *The formulation in Eq. (2) is equivalent to the following objective function:*

$$\mathcal{R}_S = tr(\hat{\mathbf{Y}}\mathcal{L}\hat{\mathbf{Y}}^T), \quad (3)$$

where the Laplacian matrix [Chung, 2005] $\mathcal{L}$ is defined as

$$\mathcal{L} = \boldsymbol{\Pi} - \frac{\boldsymbol{\Pi}\mathbf{P} + \mathbf{P}^T\boldsymbol{\Pi}}{2}. \quad (4)$$

Proof. The proof is straightforward and can be also found in previous work [Chung, 2005; Zhou *et al.*, 2005]. □

3.2 Modeling Content Information

One widely used method for text analytics is Least Squares [Lawson and Hanson, 1995], which learns a linear model to fit the training data. The classification task can be performed by solving

$$\min_{\mathbf{W}} \ \frac{1}{2}\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_F^2, \quad (5)$$

where $\mathbf{X}$ is the content matrix of training data, and $\mathbf{Y}$ is the label matrix. This formulation is to minimize the learning error between the predicted value $\hat{\mathbf{Y}} = \mathbf{X}^T\mathbf{W}$ and the true value $\mathbf{Y}$ in the training data.

As we discussed, microblogging messages are noisy and unstructured. The traditional text representation methods, like the "Bag of Words" or the N-gram model, often lead to the "curse of dimensionality." It is also observed that when people speed-read a text, they may not fully parse every word but instead seek a sparse representation with a few key phrases or words [Marinis, 2003]. In addition, by providing some meaningful words rather than non-intuitive ones, it may help sociologists, security engineers, and even the public understand the motivation and behavior of social spammers. So we are motivated to exploit sparse learning [Ghanoui *et al.*, 2011], which allows better interpretability of the learning results for social spammer detection.

Sparse learning methods have been used in various applications to obtain a more efficient and interpretable model. One of the most widely used methods is the lasso [Friedman *et al.*, 2008], which introduces an ℓ_1-norm penalization on Least Squares. The classifier can be learned by solving the following optimization problem:

$$\min_{\mathbf{W}} \ \frac{1}{2}\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_F^2 + \lambda_1\|\mathbf{W}\|_1, \quad (6)$$

where $\|\mathbf{W}\|_1 = \sum_{i=1}^{m}\sum_{j=1}^{c}|\mathbf{W}_{ij}|$, and λ_1 is the sparse regularization parameter. The second term leads to a sparse representation of the learned model.

As pointed out by Zou and Hastie [2005], if there is a group of variables among which the pairwise correlations are very high, then the lasso tends to randomly select variables from this group. To make the sparse learning more stable, we further employ elastic net [Zou and Hastie, 2005], which does automatic variable selection and continuous shrinkage, and can select groups of correlated variables. It is formulated by further adding a Frobenius norm regularization on the learned model as follows:

$$\min_{\mathbf{W}} \ \frac{1}{2}\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_F^2 + \lambda_1\|\mathbf{W}\|_1 + \frac{\lambda_2}{2}\|\mathbf{W}\|_F^2, \quad (7)$$

where λ_1 and λ_2 are positive regularization parameters to control the sparsity and robustness of the learned model.

3.3 Social Spammer Detection in Microblogging

Many existing text classification methods assume that instances are independent and identically distributed (i.i.d.). They focus on either building a sophisticated feature space or employing effective classifiers to achieve better classification performance, without taking advantage of the fact that the instances are networked with each other. In the problem of social spammer detection, microblogging users are connected via social networks. We propose to consider both network and content information in a unified model.

Since $\hat{\mathbf{Y}} = \mathbf{X}^T\mathbf{W}$, Eq. (3) can be easily rewritten as

$$\mathcal{R}_S = tr(\mathbf{W}^T\mathbf{X}\mathcal{L}\mathbf{X}^T\mathbf{W}). \quad (8)$$

By considering both network and content information, the social spammer detection can be formulated as the following optimization problem:

$$\min_{\mathbf{W}} \ \frac{1}{2}\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_F^2 + \lambda_1\|\mathbf{W}\|_1 + \frac{\lambda_2}{2}\|\mathbf{W}\|_F^2 + \frac{\lambda_s}{2}tr(\mathbf{W}^T\mathbf{X}\mathcal{L}\mathbf{X}^T\mathbf{W}). \quad (9)$$

By solving Eq. (9), the identity label of each unknown target user $\mathbf{x}$ can be predicted by

$$\arg\max_{i \in \{spammer, normal\}} \mathbf{x}^T\mathbf{w}_i. \quad (10)$$

We next introduce an efficient algorithm to solve the optimization problem in Eq. (9).

3.4 An Optimization Algorithm

The optimization problem in Eq. (9) is convex and non-smooth. Following [Liu *et al.*, 2009; Nesterov and Nesterov, 2004], the basic idea of the proposed algorithm is to reformulate the non-smooth optimization problem as an equivalent smooth convex optimization problem.

Lemma 1 $\|\mathbf{W}\|_1$ *is a valid norm.*

Proof. It is easy to verify that $\|\mathbf{W}\|_1$ satisfies the three conditions of a valid norm, including the triangle inequality $\|\mathbf{A}\|_1 + \|\mathbf{B}\|_1 \leq \|\mathbf{A} + \mathbf{B}\|_1$, which completes the proof. □

Theorem 2 *Eq. (9) can be reformulated as a constrained smooth convex optimization problem:*

$$\min_{\mathbf{W}\in\mathcal{Z}} \mathcal{O}(\mathbf{W}) = \frac{1}{2}\|\mathbf{X}^T\mathbf{W}-\mathbf{Y}\|_F^2 + \frac{\lambda_2}{2}\|\mathbf{W}\|_F^2, + \frac{\lambda_s}{2} tr(\mathbf{W}^T\mathbf{X}\mathcal{L}\mathbf{X}^T\mathbf{W}), \quad (11)$$

where

$$\mathcal{Z} = \{\mathbf{W}| \ \|\mathbf{W}\|_1 \leq z\}, \quad (12)$$

and $z \geq 0$ is the radius of the ℓ_1-ball. Note that λ_1 and z have a one-to-one correspondence between each other.

Proof. Since $\|\mathbf{W}\|_1$ is a valid norm, it defines a closed and convex set $\mathcal{Z}$. The Hessian matrix of the reformulated objective function $\mathcal{O}(\mathbf{W})$ is positive semi-definite. Thus the optimization problem in Eq. (11) is convex and differentiable. Our problem defines a convex and differentiable function $\mathcal{O}(\mathbf{W})$ in a closed and convex set $\mathcal{Z}$. Thus the reformulated function is a constrained smooth convex optimization problem, which completes the proof. □

A widely used method, proximal gradient descent [Ji and Ye, 2009], is employed to optimize the above constrained smooth convex problem. The method solves the problem by updating the following,

$$\mathbf{W}_{t+1} = \arg\min_{\mathbf{W}\in\mathcal{Z}} M_{\gamma,\mathbf{W}_t}(\mathbf{W}), \quad (13)$$

where $M_{\gamma,\mathbf{W}_t}(\mathbf{W})$ is the Euclidean projection [Boyd and Vandenberghe, 2004, Chapter 8.1], which is defined as

$$M_{\gamma,\mathbf{W}_t}(\mathbf{W}) = \mathcal{O}(\mathbf{W}_t) + \langle\nabla\mathcal{O}(\mathbf{W}_t), \mathbf{W}-\mathbf{W}_t\rangle + \frac{\gamma}{2}\|\mathbf{W}-\mathbf{W}_t\|_F^2, \quad (14)$$

where γ is the step size, and

$$\nabla\mathcal{O}(\mathbf{W}_t) = \mathbf{X}\mathbf{X}^T\mathbf{W}_t - \mathbf{X}\mathbf{Y} + \lambda_2\mathbf{W} + \lambda_s\mathbf{X}\mathcal{L}\mathbf{X}^T\mathbf{W}_t. \quad (15)$$

Let $\mathbf{U}_t = \mathbf{W}_t - \frac{1}{\gamma}\nabla\mathcal{O}(\mathbf{W}_t)$. The Euclidean projection has a closed-form solution [Liu *et al.*, 2009] as follows:

$$\mathbf{w}_{t+1}^j = \begin{cases} (1-\frac{\lambda_1}{\gamma\|\mathbf{u}_t^j\|})\mathbf{u}_t^j & \text{if } \|\mathbf{u}_t^j\| \geq \frac{\lambda_1}{\gamma} \\ 0 \quad \text{otherwise} \end{cases} \quad (16)$$

where $\mathbf{u}_t^j$, $\mathbf{w}^j$ and $\mathbf{w}_t^j$ are the j-th rows of $\mathbf{U}_t$, $\mathbf{W}$ and $\mathbf{W}_t$, respectively. We now introduce the detailed algorithm of *SSDM* in Algorithm 1.

In the algorithm, we use Nesterov's method [Nesterov and Nesterov, 2004] to solve the optimization problem in Eq. (9) from line 2 to 13. To accelerate the proximal gradient descent in Eq. (13), we construct a linear combination [Ji and Ye, 2009] of $\mathbf{W}_t$ and $\mathbf{W}_{t-1}$ to update $\mathbf{H}_t$ in line 3. Lines 4 to 11 correspond to the line search algorithm for γ according to the Armijo-Goldstein rule. Based on this algorithm, we can have an efficient and optimal solution to the convex optimization problem. Similar to the proof in [Liu *et al.*, 2009], it is easy to verify that the convergence rate of the proposed algorithm is $O(\frac{1}{\sqrt{\epsilon}})$ for achieving an accuracy of ϵ.

Algorithm 1: SSDM – Social Spammer Detection in Microblogging

Input: $\{\mathbf{X}, \mathbf{Y}, \mathcal{L}, \mathbf{W}_0, \lambda_1, \lambda_2, \lambda_s\}$
Output: $\mathbf{W}$

1: Initialize $\eta_0 = 0$, $\eta_1 = 1$, $\mathbf{W}_1 = \mathbf{W}_0$, $t = 1$
2: **repeat**
3: Set $\mathbf{H}_t = \mathbf{W}_t + \frac{\eta_{t-1}-1}{\eta_t}(\mathbf{W}_t - \mathbf{W}_{t-1})$
4: **loop**
5: Set $\mathbf{U}_t = \mathbf{H}_t - \frac{1}{\gamma}\nabla\mathcal{O}(\mathbf{W}_t)$
6: Compute $\mathbf{W}_{t+1}$ according to Eq. (16)
7: **if** $\mathcal{O}(\mathbf{W}_{t+1}) \leq M_{\gamma,\mathbf{H}_t}(\mathbf{W}_{t+1})$ **then**
8: $\lambda_{t+1} = \gamma$, break
9: **end if**
10: $\gamma = 2 \times \gamma$
11: **end loop**
12: $\mathbf{W} = \mathbf{W_{t+1}}$, $\eta_{t+1} = \frac{1+\sqrt{1+4\eta_t}}{2}$
13: **until** convergence

4 Experiments

In this section, we conduct experiments to assess the effectiveness of the proposed framework *SSDM*. Through the experiments, we aim to answer the following two questions:

1. How effective is the proposed framework compared with other methods of social spammer detection?
2. What are the effects of the social network and content information on the social spammer detection?

We begin by introducing the dataset and experimental setup and then compare the performance of different spammer detection methods. Finally, we study the effects of important parameters on the proposed method.

4.1 Dataset

We now introduce the real-world Twitter dataset used in our experiment. A data crawling process, which is similar to [Thomas *et al.*, 2011; Yang *et al.*, 2011; Zhu *et al.*, 2012], is employed to construct the dataset. We first crawled a Twitter dataset from July 2012 to September 2012 via the Twitter Search API.[2] The users that were suspended by Twitter during this period are considered as the gold standard [Thomas *et al.*, 2011] of spammers in the experiment. We then randomly sampled the normal users which have social relations with the spammers. According to the literature of spammer detection, the two classes are imbalanced, i.e., the number of normal users we sampled is much greater than that of spammers in the dataset. We finally remove stop-words and perform stemming for all the tweets. The statistics of the dataset is presented in Table 2.

4.2 Experimental Setup

We follow standard experiment settings used in [Benevenuto *et al.*, 2010; Zhu *et al.*, 2012] to evaluate the performance of spammer detection methods. In particular, we apply different methods on the Twitter dataset. Precision, recall, and F_1-measure are used as the performance metrics.

[2] https://dev.twitter.com/docs/api/1/get/search/

Table 1: Social Spammer Detection Results

	50% of the Training Data			*100% of the Training Data*		
	Precision	Recall	F_1-measure (gain)	Precision	Recall	F_1-measure (gain)
LS_Content_SN	0.786	0.843	0.813 (N.A.)	0.793	0.850	0.821 (N.A.)
EN_Content_SN	0.801	0.872	0.835 (+2.69%)	0.836	0.891	0.863 (+5.09%)
SMF_UniSN	0.804	0.889	0.845 (+3.87%)	0.844	0.915	0.878 (+6.92%)
SSDM	0.852	0.896	0.873 (+7.40%)	0.865	0.939	0.901 (+9.73%)

Table 2: Summary of the Experimental Dataset

# Spammers	# Normal Users	Max Degree of Users
2,118	10,335	1,025
# Tweets	# Unigrams	Min Degree of Users
380,799	21,388	3

There are three positive parameters involved in the experiments, including λ_1, λ_2, and λ_s in Eq. (9). λ_1 is to control the sparsity of the learned model, λ_2 is the parameter to make the learned model more robust, and λ_s is to control the contribution of network information. As a common practice, all the parameters can be tuned via cross-validation with validation data. In the experiments, we empirically set $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, and $\lambda_s = 0.1$ for general experiment purposes. The effects of the parameters on the learning model will be further discussed in Section 4.5.

4.3 Performance Evaluation

To answer the first question asked in the beginning of Section 4, we compare our proposed method *SSDM* with the following baseline methods. All the methods utilize both content and network information in different ways.

- *LS_Content_SN*: the Least Squares [Lawson and Hanson, 1995] is a widely used classifier for i.i.d. data. We combine the content matrix $\mathbf{X}$ and adjacency matrix $\mathbf{G}$ of the social network together for user representation.
- *EN_Content_SN*: the elastic net is one of the most effective sparse learning methods [Zou and Hastie, 2005], and it is applied on the same data matrix as the first baseline.
- *SMF_UniSN*: a multi-label informed latent semantic indexing [Yu *et al.*, 2005; Zhu *et al.*, 2012] is used to model the content information, and undirected graph Laplacian [Chung, 1997] is used to incorporate the network information. This is the state-of-the-art method for spammer detection in an undirected social network. In the experiment, we convert the directed graph to an undirected one with $\mathbf{G} = max(\mathbf{G}, \mathbf{G}^T)$.
- *SSDM*: our proposed method for spammer detection.

The experimental results of the methods are presented in Table 1. In the experiment, we use five-fold cross validation for all the methods. To avoid effects brought by the size of the training data, we conduct two sets of experiments with different numbers of training samples. In each round of the experiment, 80% of the whole dataset is held for training. In the table, "50% of Training Data" means that we randomly chose 50% of the 80%, thus using 40% of the whole dataset for training. Also, "gain" represents the percentage improvement of the methods in comparison with our first baseline method *LS_Content_SN*. In the experiment, each result denotes an average of 10 test runs. By comparing the results of different methods, we draw the following observations:

(1) From the results in Table 1, we can observe that our proposed framework *SSDM* consistently outperforms other baseline methods using all metrics with different sizes of training data. Our method achieves better performance than the state-of-the-art method *SMF_UniSN*. We apply two-sample one-tail t-tests to compare *SSDM* with the three baseline methods. The experiment results demonstrate that *SSDM* performs significantly better (with the significance level $\alpha = 0.01$) than the three methods. This indicates that, compared with other methods, our proposed model successfully utilizes both content and network information for social spammer detection.

(2) Among the three baseline methods, *LS_Content_SN* achieves the worst performance. With the introduction of sparsity regularization, *EN_Content_SN* has performance improvement. This demonstrates that sparse learning is effective to handle the noisy and high-dimensional data in microblogging. *SMF_UniSN* achieves the best performance, which indicates that simply combining content and network features together does not work well in the first two methods.

In summary, the methods perform differently in social spammer detection. In most cases, the simple combination of content and network information does not work well. It suggests that the way of using the two kinds of information is important. The superior performance of our proposed method answers the first question that, compared with other methods, *SSDM* is effective in social spammer detection.

4.4 Effects of Network and Content Information

In this subsection, we study the importance of each kind of information and accordingly answer the second question asked in the beginning of this section. We compare our proposed method with the following two groups of four methods:

- *Content-based methods*: support vector machine (SVM) and elastic net (EN) are employed for spammer detection based on content information only.
- *Network-based methods*: SVM and EN are employed based on network information, which is represented as the adjacency matrix of the social network.

We compare the performance of our proposed framework *SSDM* and the methods with only one type of information on

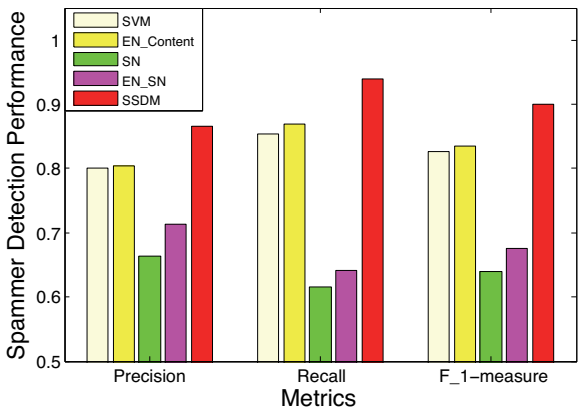

Figure 1: Social Spammer Detection Performance

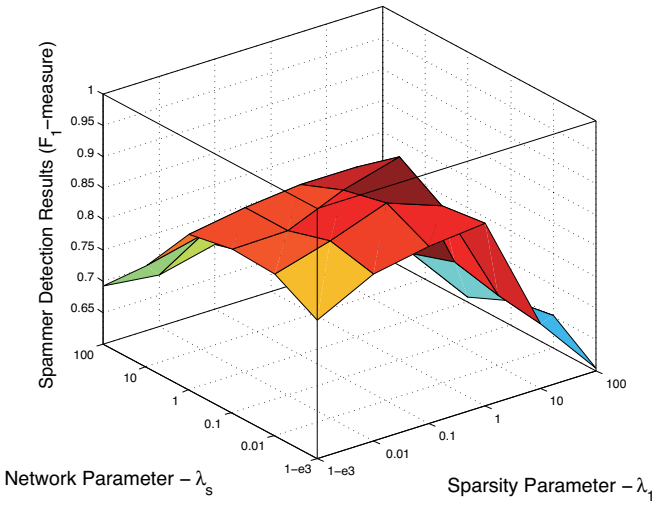

Figure 2: Impact of the Sparsity Parameter (λ_1) and the Network Parameter (λ_s)

the Twitter dataset. The results are plotted in Figure 1. The first four bars represent the performance of the two representative methods *SVM* and *EN* with one type of information, respectively. The last is our proposed method *SSDM*.

From the figure, we observe that, with the integration of content and network information in a unified model, the proposed framework *SSDM* achieves better performance than those with only one kind of information. Among the four baseline methods, *SVM_Content* and *EN_Content* have comparable performance. They significantly outperform the other two methods *SVM_SN* and *EN_SN*. This demonstrates that, in this experiment, content information is more effective than social network information. We need a more sophisticated way to represent social network information for social spammer detection. Simply employing neighbors of a user for representation does not work well.

In summary, the methods based on network information do not have good performance in social spammer detection. It suggests that the way of integrating social network information is important. The superior performance of our proposed method *SSDM* further validates its excellent use of both network and content information in a unified way.

4.5 Parameter Analysis

As discussed in Section 4.2, the effects of two important parameters, i.e., λ_1 and λ_s, need to be further explored. λ_1 is to control the sparseness of the learned model, and λ_s is to control the contribution of social network information to the model. We now conduct experiments to compare the social spammer detection performance of the proposed *SSDM* on the Twitter dataset with different parameter settings.

The social spammer detection results (F_1-measure) of *SSDM* with different parameter settings on the dataset are plotted in Figure 2. In the figure, performance of *SSDM* improves as the parameters λ_1 and λ_s increase, and reaches a peak at $\lambda_1 = 0.1$ and $\lambda_s = 1$. When $\lambda_1 > 0.1$ or $\lambda_s > 1$, the performance of *SSDM* declines. Generally, the performance is not very sensitive to λ_1 when it is in a reasonable range [0.01, 10]. The performance changes significantly when $\lambda_s > 1$. The results suggest that the proposed framework can achieve relatively good performance when the parameters are in the range [0.01, 1].

5 Conclusion and Future Work

New features of microblogging services present great challenges to the problem of social spammer detection. In this paper, we investigate how to seamlessly integrate the network and content information of microblogging users to perform effective social spammer detection. In particular, the proposed framework models both types of information in a unified way. Also, we present an efficient algorithm to solve the proposed non-smooth convex optimization problem. Experiments on a real Twitter dataset show that our *SSDM* framework can effectively integrate both kinds of information to outperform the state-of-the-art methods.

There are many potential future extensions of this work. It would be interesting to investigate other social activities, like retweet behavior and emotion status, for social spammer detection. Also, sparse learning can generate a number of important textual features with the model. Conducting behavior and linguistic analysis across social media sites [Hu *et al.*, 2013] to better understand motivations of the social spammers with the textual features is also a promising direction.

Acknowledgments

We truly thank the anonymous reviewers for their pertinent comments. This work is, in part, supported by ONR (N000141110527) and (N000141010091), and NSF (IIS-1217466).

References

[Benevenuto *et al.*, 2010] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Proceedings of CEAS*, 2010.

[Bilge *et al.*, 2009] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your contacts are belong to us: automated identity theft attacks on social networks. In *Proceedings of WWW*, 2009.

[Boyd and Vandenberghe, 2004] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

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

[Brown *et al.*, 2008] G. Brown, T. Howe, M. Ihbe, A. Prakash, and K. Borders. Social networks and context-aware spam. In *Proceedings of CSCW*, 2008.

[Chung, 1997] F.R.K. Chung. *Spectral graph theory*. Number 92. Amer Mathematical Society, 1997.

[Chung, 2005] F. Chung. Laplacians and the cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.

[Danezis and Mittal, 2009] G. Danezis and P. Mittal. Sybilinfer: Detecting sybil nodes using social networks. 2009.

[Friedman *et al.*, 2008] J. Friedman, T. Hastie, and R. Tibshirani. The elements of statistical learning, 2008.

[Ghanoui *et al.*, 2011] L. Ghanoui, G. Li, V. Duong, V. Pham, A. Srivastava, and K. Bhaduri. Sparse machine learning methods for understanding large text corpora. In *Proceedings of CIDU*, 2011.

[Ghosh *et al.*, 2012] 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 WWW*, 2012.

[Grier *et al.*, 2010] 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.

[Gu and Han, 2011] Quanquan Gu and Jiawei Han. Towards feature selection in network. In *Proceedings of CIKM*, 2011.

[Hu *et al.*, 2009] Xia Hu, Nan Sun, Chao Zhang, and Tat-Seng Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of CIKM*, 2009.

[Hu *et al.*, 2013] Yuheng Hu, Kartik Talamadupula, and Subbarao Kambhampati. Dude, srsly?: The surprisingly formal nature of twitters language. *Proceedings of ICWSM*, 2013.

[Ji and Ye, 2009] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of ICML*, 2009.

[Lawson and Hanson, 1995] C.L. Lawson and R.J. Hanson. *Solving least squares problems*, volume 15. SIAM, 1995.

[Lee *et al.*, 2010] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots + machine learning. In *Proceedings of SIGIR*, 2010.

[Liu *et al.*, 2009] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of UAI*, 2009.

[Marinis, 2003] T. Marinis. Psycholinguistic techniques in second language acquisition research. *Second Language Research*, 19(2):144–161, 2003.

[Nesterov and Nesterov, 2004] Y. Nesterov and I.U.E. Nesterov. *Introductory lectures on convex optimization: A basic course*. 2004.

[O'Callaghan *et al.*, 2012] Derek O'Callaghan, Martin Harrigan, Joe Carthy, and Pdraig Cunningham. Network analysis of recurring youtube spam campaigns. In *Proceedings of ICWSM*, 2012.

[Ratkiewicz *et al.*, 2011a] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *Proceedings of ICWSM*, 2011.

[Ratkiewicz *et al.*, 2011b] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of WWW*, 2011.

[Thomas *et al.*, 2011] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of ACM SIGCOMM conference on Internet measurement conference*, 2011.

[Webb *et al.*, 2008] S. Webb, J. Caverlee, and C. Pu. Social honeypots: Making friends with a spammer near you. In *Proceedings of CEAS*, 2008.

[Weng *et al.*, 2010] J. Weng, E.P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of WSDM*, 2010.

[Yang *et al.*, 2011] Z. Yang, C. Wilson, X. Wang, T. Gao, B.Y. Zhao, and Y. Dai. Uncovering social network sybils in the wild. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 259–268. ACM, 2011.

[Yu *et al.*, 2005] K. Yu, S. Yu, and V. Tresp. Multi-label informed latent semantic indexing. In *Proceedings of SIGIR*, 2005.

[Zhou *et al.*, 2005] D. Zhou, J. Huang, and B. Schölkopf. Learning from labeled and unlabeled data on a directed graph. In *Proceedings of ICML*, 2005.

[Zhu *et al.*, 2012] Y. Zhu, X. Wang, E. Zhong, N.N. Liu, H. Li, and Q. Yang. Discovering spammers in social networks. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[Zou and Hastie, 2005] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.