Social Spammer Detection with Sentiment Information

Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu
Computer Science and Engineering
Arizona State University, Tempe, AZ 85287, USA
Email: {xia.hu, jiliang.tang, huiji.gao, huan.liu}@asu.edu

***Abstract*—Social media is a popular platform for spammers to unfairly overwhelm normal users with unwanted or fake content via social networking. The spammers significantly hinder the use of social media systems for effective information dissemination and sharing. Different from the spammers in traditional platforms such as email and the Web, spammers in social media can easily connect with each other, sometimes without mutual consent. They collude with each other to imitate normal users by quickly accumulating a large number of "human" friends. In addition, content information in social media is noisy and unstructured. It is infeasible to directly apply traditional spammer detection methods in social media. Understanding and detecting deception has been extensively studied in traditional sociology and social sciences. Motivated by psychological findings in physical world, we investigate whether sentiment analysis can help spammer detection in online social media. In particular, we first conduct an exploratory study to analyze the sentiment differences between spammers and normal users; and then present an optimization formulation that incorporates sentiment information into a novel social spammer detection framework. Experimental results on real-world social media datasets show the superior performance of the proposed framework by harnessing sentiment analysis for social spammer detection.**

I. Introduction

Social media services, like Twitter and Facebook, have become more and more popular in various scenarios such as marketing, journalism or public relations. With the growing availability of social media, social spammers [1] have emerged to unfairly overwhelm normal users with unwanted or fake content via social networking. Social spammers can be coordinated to launch various attacks such as befriending victims and then grabbing their personal information [2], conducting spam campaigns which lead to phishing, malware, and scams [3], and conducting political astroturf [4], [5]. The spamming in social media significantly hinders the quality of social networking for effective information dissemination and sharing. Successful social spammer detection is important to improve the quality of user experience, and to positively impact the overall value of the social media systems [6].

Spammer detection has been studied for years in traditional platforms such as email and the Web, which differ substantially from social media services. First, social media services allow users to easily connect with each other, sometimes without mutual consent. For example, in Twitter, 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 [7]. This reflexive reciprocity makes it easier for social spammers to collude with each other to imitate normal users by quickly accumulating a large number of social relations. Second, content information in social media is noisy and unstructured. 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 "it's cooool" are popular in social media, 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. The characteristics of social media services present great challenges to capture the deception of social spammers.

Understanding and detecting deception has been extensively studied in psychology and social sciences. It is well-established that *microexpressions* [8] play a distinct role in detecting deception. Microexpression is an involuntary facial expression of humans according to sentiments experienced. It usually occurs when a person is consciously trying to conceal all signs of how he or she is feeling [8]. Ekman [9] reported that facial and emotional "microexpressions" could be useful to assist in lie detection after testing a total of 20,000 people [10] from all walks of life. Also, as pointed out by Matsumoto *et al.* [11], one may not conclude that someone is lying if a microexpression is detected but that there is more to the story than is being told. Inspired by the psychological findings, we explore whether the utilization of sentiment information could help capture deceptions of the social spammers.

In this paper, we study the problem of utilizing sentiment information for effective social spammer detection. Specifically, we investigate the following questions: Is sentiment information potentially useful for social spammer detection? How can sentiment information be explicitly represented and incorporated for social spammer detection? Is the integration of sentiment analysis helpful for our studied problem? To answer these questions, it results in a novel framework for social Spammer Detection with Sentiment information (*SDS*). In particular, we first investigate whether sentiment differences between spammers and normal users exist in social media data. Then we discuss how to model sentiment information, combined with content and network information, in a novel social spammer detection framework. Finally, we conduct extensive experiments to evaluate the proposed model. The main contributions of the paper are outlined as follows,

- Formally define the problem of social spammer detection with sentiment information;

[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.

- Verify the sentiment differences between spammers and normal users with hypothesis testing, and model the sentiment information for spammer detection;
- Present a novel framework to incorporate sentiment information, combined with content and network information, for social spammer detection; and
- Empirically evaluate the proposed method on real-world Twitter datasets and elaborate the effects of sentiment information on our studied problem.

The remainder of this paper is structured as follows. In Section II, we define the problem of social spammer detection with sentiment information. In Section III, we conduct an exploratory study to examine the potential impacts of sentiment information for the problem. In Section IV, we propose a novel social spammer detection framework that considers sentiment, content and network information. In Section V, we conduct experiments on Twitter datasets to evaluate the proposed framework. In Section VI, we review related work. In Section VII, we conclude and present the future work.

II. Problem Statement

One distinct feature of social media data is that it provides abundant contextual information other than social networks. The problem we studied is different from traditional spammer detection in social networks since the latter typically only considers either the content or network information [12], [13]. In this section, we first present the notation used in this paper and then formally define the problem of social spammer detection with sentiment information.

Notation: lowercase bold letters (e.g., $\mathbf{a}$) denote column vectors, upper-case letters (e.g., $\mathbf{A}$) denote matrices, and lower-case letters (e.g., a) denote scalars. $\mathbf{A}(i,j)$ denotes the entry at the i^{th} row and j^{th} column of a matrix $\mathbf{A}$. Let $\|\mathbf{A}\|$ denote the Euclidean norm, and $\|\mathbf{A}\|_F$ the Frobenius norm of the matrix $\mathbf{A}$. Specifically, $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{A}(i,j)^2}$. Let $\mathbf{A}^T$ and $Tr(\mathbf{A})$ denote the transpose and trace of $\mathbf{A}$, respectively.

Let $\mathbf{S} = [\mathbf{X}, \mathcal{G}, \mathbf{Y}]$ be a target user set with content information $\mathbf{X}$, social network information $\mathcal{G}$ and identity label matrix $\mathbf{Y}$. We use user-word matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ to denote content information, i.e., posts written by the users, where n is the number of users, and m is the number of textual features. We use $\mathcal{G} = (V, E)$ to denote the social network, where nodes $v \in V$ represent social media users, and each directed edge between two nodes $[u, v] \in E$ represents a following relation from u to v. We do not have self-links in the graph, i.e., $u \neq v$. $\mathbf{Y} \in \mathbb{R}^{n\times c}$ denotes the identity label matrix, where c is the number of possible identity labels. In this paper, we focus on the binary classification problem, i.e., $c = 2$ and the users will be classified as spammers or normal users. It is practical to extend this setting to a multi-class classification task.

Given another corpus of posts $\mathbf{C} \in \mathbb{R}^{t\times m}$ with sentiment labels, where t is the number of posts, and m is the number of textual features. We use $\mathbf{s} \in [-1, 1]^t$ to represent the sentiment polarity labels of the corresponding social media posts. For example, $\mathbf{s}(i) = 1$ represents that sentiment of the i^{th} post in the corpus is positive, and $\mathbf{s}(i) = -1$ negative.

We now formally define the problem as follows:

Given a set of social media users $\mathbf{S}$ with content information $\mathbf{X}$, social network information $\mathcal{G}$, and identity label information $\mathbf{Y}$ of part of the users in the set (i.e., training data), we can also learn the sentiment information from another set of labeled posts $[\mathbf{C}, \mathbf{s}]$, our goal is to learn a model to automatically assign identity labels for unknown users (i.e., test data) as spammers or normal users.

TABLE I. Statistics of the Datasets

Statistics	*TUSH*	*TSS*
# of Spammers	16,841	4,005
# of Normal Users	13,697	15,832
# of Unigrams	31,004	18,055

III. Data and Exploratory Study

A major motivation of this study is to investigate if sentiment information is useful for social spammer detection. Before proceeding further, we first introduce real-world datasets used in this work and examine whether sentiment information has any potential impact for social spammer detection.

A. Datasets

Three Twitter datasets are used in our study. The first two contain labels for social spammer detection, i.e., TAMU Social Honeypots and Twitter Suspended Spammers, and the third one Stanford Twitter Sentiment has sentiment labels. Now we introduce the three datasets in detail.

TAMU Social Honeypots Dataset (TUSH):[2] Lee *et al.* [14] created a collection of 41,499 Twitter users with identity labels as spammers and normal users. The dataset was collected from December 30, 2009 to August 2, 2010 on Twitter. It consists of users, their number of followers and posted tweets. We further refined the dataset according to users' social relation information, which is a complete follower graph[3] crawled by Kwak *et al.* [15] during July 2009. According to the social network, we filter the users who post less than two tweets or have less than two friends in the dataset. Finally, it leaves a corpus of 30,538 users that consists of 16,841 spammers and 13,697 normal users. This dataset has balanced number of spammers and normal users.

Twitter Suspended Spammers Dataset (TSS): We used a data construction process, which is similar to [16], [17], to build this dataset. We first crawled a Twitter dataset from August 5, 2013 to October 11, 2013 using the Twitter Search API.[4] We examined all of the crawled users at the end of the crawling process. The users that were suspended by Twitter during this period are considered as the gold standard [17] of spammers in the experiment. We then randomly sampled normal users which have social relations with the spammers. To consider effects brought by different class distribution, according to the literature of social spammer detection [6], we made the two classes in TSS imbalanced, i.e., the number of normal users we sampled is much greater than that of spammers in the dataset. In addition, users that post less than

[2] http://infolab.tamu.edu/data/

[3] http://an.kaist.ac.kr/traces/WWW2010.html/

[4] http://dev.twitter.com/docs/api/

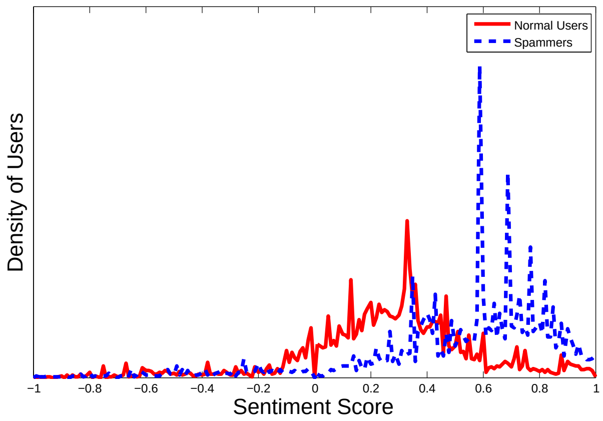

Fig. 1. Sentiment Score Distribution

two tweets or have less than two friends in the whole dataset are removed. Finally, it leaves a corpus of 19,837 users that consists of 4,005 spammers and 15,832 normal users.

A standard procedure is used for data preprocessing on both datasets. All of the non-English tweets are filtered out from the datasets. We remove stop-words and perform stemming for all the tweets. The unigram model is employed to construct the feature space, tf-idf is used as the feature weight. The statistics of the datasets are presented in Table I.

Stanford Twitter Sentiment (SENT)[5]: Go *et al.* [18] created a collection of 40,216 tweets with polarity sentiment labels to train a sentiment classifier. The tweets in the dataset are crawled between April 6, 2009 and June 25, 2009. All the tweets and corresponding sentiment labels in the dataset are used to learn a model for sentiment analysis.

B. Sentiment Distribution

We employ a standard method to compute the sentiment score of each user. In particular, a supervised sentiment analysis model is learned based on the labeled dataset SENT, and we then apply the learned model to compute the sentiment score of users in the two datasets TUSH and TSS.

Pang and Lee [19] conducted experiments to study the effectiveness of different methods on sentiment analysis. It shows that machine learning techniques can achieve good performance on benchmark datasets. Following widely used sentiment analysis methods introduced in [19], [20], [21], a linear regression [22] is employed to fit the learned model to sentiment labels $\mathbf{s}$. The linear regression aims to learn a model by solving the following optimization problem:

$$\min_{\mathbf{w}} \quad \|\mathbf{C}\mathbf{w} - \mathbf{s}\|^2, \tag{1}$$

where $\mathbf{C}$ represents the content matrix of SENT dataset, $\mathbf{w}$ represents the learned coefficients of the features, and $\mathbf{s}$ denotes the sentiment labels of the posts in $\mathbf{C}$. This formulation is a traditional supervised learning method, and it has a closed-form solution: $\mathbf{w} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{s}$. By solving Eq. (1), the sentiment score of a user u can be computed by $\mathbf{X}(u)\mathbf{w}$.

TABLE II. T-Test Results (P-Values) to Verify Microexpressions

	TUSH	TSS
Microexpressions	<0.938e-9	<1.011e-15

We compute the sentiment score of each user in the two datasets. The sentiment scores are normalized in the range of $[-1, 1]$. We plot the polarity score distributions of spammers and normal users on the TUSH dataset in Figure 1. In the figure, x axis represents the sentiment score and y axis the density of users who have the exact sentiment score. Red curve denotes the sentiment score distribution of normal users and blue dotted curve the distribution of spammers. From the figure, we can observe two normal-like distributions for spammers and normal users. The two distributions center with different mean values and show clearly different patterns. It suggests that the sentiment patterns of normal users and spammers are different. Similar results have been observed on the TSS dataset; we omit the results owing to lack of space.

C. Verifying Sentiment Correlation

The preliminary results in Section III-B show that the sentiment distributions of spammers and normal users are different. We now further verify whether this observation is potential useful for our studied problem.

The psychological finding of microexpression suggests that sentiments of spammers are different from normal users. The assumption is that the sentiments of two users with the same identity, i.e., both are spammers or normal users, have higher probability to be consistent than those of two random users. We use hypothesis testing to validate whether this assumption of sentiment consistency holds in the two Twitter datasets.

We first define the sentiment difference score $d(i, j)$ between two users as

$$d(i, j) = ||\mathbf{s}(i) - \mathbf{s}(j)||_2, \tag{2}$$

where $\mathbf{s}(i)$ and $\mathbf{s}(j)$ represent sentiment scores of the two users. The sentiment scores are computed by the method we introduced in Section III-B.

Then, two vectors $\mathbf{s}_c$ and $\mathbf{s}_r$ with an equal number of elements are constructed. Each element of the first vector $\mathbf{s}_c$ is calculated by Eq. (2), where $\mathbf{s}(i)$ and $\mathbf{s}(j)$ are users with the same identity. Each element of the second vector represents the sentient difference score between $\mathbf{s}(i)$ and $\mathbf{s}(r)$, which denotes the sentiment score of another randomly selected user. We form a two-sample one-tail t-test to validate the assumption. We test whether there is sufficient evidence to support the hypothesis that sentiment difference of the first group is greater or equal than that of the second. The null hypothesis and alternative hypothesis are formulated as follows:

$$\begin{aligned} H_0 &: \mu_c - \mu_r \geq 0 \\ H_1 &: \mu_c - \mu_r < 0 \end{aligned} \tag{3}$$

where μ_c and μ_r represent the sample means of sentiment difference scores in the two groups, respectively.

The t-test results, p-values, are summarized in Table II. The results suggest that there is strong statistical evidence, with significance level $\alpha = 0.01$, to reject the null hypothesis on

[5] http://www.stanford.edu/~alecmgo/cs224n/

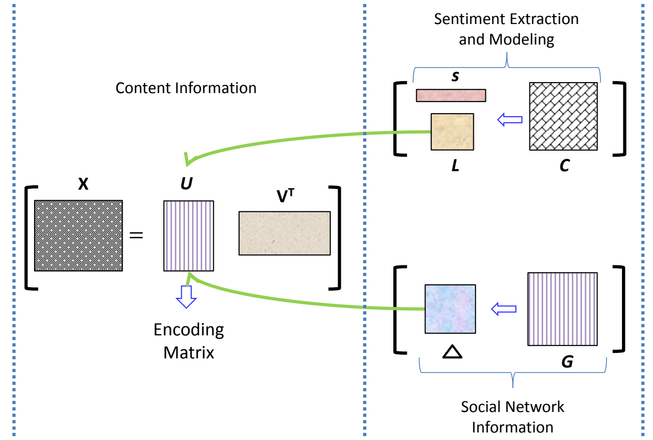

Fig. 2. Illustration of the Spammer Detection Framework

the two datasets. In other words, we validate the assumption in the two datasets. This exploratory study paves the way for our next step: how to explicitly model and utilize the sentiment information for social spammer detection.

IV. SDS: Social Spammer Detection with Sentiment Information

In this section, we introduce the proposed framework that incorporates sentiment, content and social network information for social spammer detection in detail.

We plot the work flow of our proposed framework in Figure 2. From the figure, we can see that the whole framework consists of three components. The left part represents modeling of content information. There are two constraints on the learned factor matrix **U** which is derived from content information. As shown in the upper right part of the figure, the first constraint is from sentiment information $\mathcal{L}$, which is learned from an independent sentiment related source **C**. As shown in the lower right part of the figure, the second constraint is learned from social network information **G**. In this section, we first discuss how to model content information, and then introduce the modeling of sentiment and network information to detect social spammers. Finally, we present the framework that considers the three types of information as well as its computational algorithm for social spammer detection.

A. Modeling Content Information

Social media provides abundant content information. Unlike spam detection in platforms such as email and SMS, content analysis has been little studied for social spammer detection. To make use of content information, a straightforward way is to learn a supervised model based on labeled data, and apply the learned model for spammer detection. However, this method yields two problems due to the unstructured and noisy content information in social media. First, text representation models, like n-gram model, often lead to a high-dimensional feature space because of the large size of data and vocabulary. Second, in addition to the short form of texts, abbreviations and acronyms are widely used in social media, thus making the data representation very sparse [23]. These distinct characteristics of social media data make traditional text analytics less applicable for our task.

To tackle the problems, we propose to model the content information from topic-level instead of learning word-level knowledge. Motivated by previous work on topic modeling [24], a user's posts usually focus on a few topics, resulting in **X** very sparse and low-rank. The proposed method is built on a non-negative matrix factorization model (NMF) [25]. NMF is to seek a more compact but accurate low-rank representation of the users by solving the following optimization problem:

$$\min_{\mathbf{U},\mathbf{V}\geq 0} \quad \|\mathbf{X}-\mathbf{U}\mathbf{V}^T\|_F^2, \tag{4}$$

where **X** is the content matrix, $\mathbf{U} \in \mathbb{R}^{n\times r}$ with $r \ll m$ is an encoding matrix that indicates a low-rank user representation in a topic space and $\mathbf{V} \in \mathbb{R}^{m\times r}$ is a mixing matrix. Both **U** and **V** are non-negative factor matrices to be learned.

The matrix factorization [26], [27] based content modeling has several nice properties: (1) this model has a nice probabilistic interpretation with Gaussian noise; (2) many existing optimization methods can be used to provide a well-worked optimal solution; (3) it can be scaled to a large number of users, which is a common setting in social media; (4) this formulation is flexible and allows us to introduce prior knowledge such as sentiment information and social network information introduced in next subsections.

B. Modeling Sentiment Information

The observation introduced in Section III suggests that the sentiments of two users with the same identity label have higher probability to be consistent. Based on this observation, we propose to model the sentiment information with graph Laplacian [28]. We construct an undirected graph $\mathbf{G}_S$ based on sentiment information of the users. In the graph, each node represents a user and each edge represents the sentiment correlation between two users. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ of the constructed graph $\mathbf{G}_S$ is formulated as the following:

$$\mathbf{A}(i,j) = \begin{cases} 1 & \text{if } u_i \in \mathcal{N}(u_j) \ or \ u_j \in \mathcal{N}(u_i) \\ 0 & \text{otherwise .} \end{cases} \tag{5}$$

where u_i and u_j are nodes, and $\mathcal{N}(u_i)$ represents the k-nearest neighbor of the user u_i in terms of sentiment information. As we discussed in Section III-B, a model **w** can be learned by minimizing the objective function in Eq. (1), and sentiment score of a user u can be computed as $\mathbf{X}(u)\mathbf{w}$. It is noted that our study is not confined to any specific sentiment analysis tools. It is practical to employ other sentiment analysis methods, e.g., lexicon-based method [29], to compute the sentiment score of each user. Since we aim to model the mutual sentiment correlation between two users, the adjacency matrix in the formulation is symmetric.

The key idea of utilizing graph Laplacian to model the sentiment information is that if two nodes are close in the graph, i.e., their sentiment scores are close to each other, the representations of the two users should be similar. It can be formulated as minimizing the following loss function:

$$\mathcal{R}_S = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mathbf{U}_i-\mathbf{U}_j)^2\mathbf{A}(i,j) \ , \tag{6}$$

where n is the number of users in the graph, $\mathbf{U}_i$ denotes representation of the i^{th} user, and $\mathbf{U}_j$ the j^{th} user. This loss function will incur a penalty if two users have different representations when they are close to each other in the constructed graph.

Let $\mathbf{D} \in \mathbb{R}^{n \times n}$ denote a diagonal matrix, and its diagonal element is the degree of a user in the adjacency matrix $\mathbf{A}$, i.e., $\mathbf{D}(i,i) = \sum_{j=1}^{n} \mathbf{A}(i,j)$.

It is easy to verify that the formulation in Eq. (6) can be rewritten as:

$$\begin{aligned}\mathcal{R}_S = \; & \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{U}_i \mathbf{A}(i,j)\mathbf{U}_i^T - \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{U}_i \mathbf{A}(i,j)\mathbf{U}_j^T \\ = \; & \sum_{i=1}^{n} \mathbf{U}_i \mathbf{D}(i,i)\mathbf{U}_i^T - \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{U}_i \mathbf{A}(i,j)\mathbf{U}_j^T \\ = \; & Tr(\mathbf{U}^T(\mathbf{D}-\mathbf{A})\mathbf{U}) \\ = \; & Tr(\mathbf{U}^T \mathcal{L}\mathbf{U}). \qquad (7)\end{aligned}$$

Besides sentiment information, abundant social network information is available in social media for social spammer detection. Next, we introduce how to model the social network information for our studied problem.

C. Modeling Social Network Information

Many efforts have been devoted to model social network information in various applications such as recommender systems [26] and trust prediction [27]. Existing methods often assume that representations of two nodes are close when they are connected with each other in the network [17], [28]. This assumption does not hold in many social media services. For example, some social media services such as microblogging allow directed following relations between users without mutual consent. In addition, as we discussed, spammers can easily follow a large number of normal users within a short time. The characteristics of the social media data make existing methods not suitable to our task.

We propose to use a variant of directed graph Laplacian to model network information. Given the social network information $\mathcal{G}$ and the identity labels $\mathbf{Y}$, four kinds of following relations can be extracted: [spammer, spammer], [normal, spammer], [normal, normal], and [spammer, normal]. Since the fourth relation [spammer, normal] can be easily faked by spammers, we only make use of the first three relations in the proposed framework. Note that this is a general setting in different social networks. In undirected social networks, e.g., Facebook, it is easy to convert the undirected graph into a direct setting. Now we introduce how to represent and model the social network information.

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

$$\mathbf{G}(i,j) = \begin{cases} 1 & \text{if } [u_i,\, u_j] \text{ is among the first three relations} \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

where u_i and u_j represent the i^{th} and j^{th} users, and $[u_i, u_j]$ is a directed edge in the graph $\mathcal{G}$.

In the social network, in-degree of the node u_i is defined as $\mathbf{d}_i^{in} = \sum_{[u_j,u_i]} \mathbf{G}(j,i)$, and out-degree of the node u is defined as $\mathbf{d}_i^{out} = \sum_{[u_i,u_j]} \mathbf{G}(i,j)$. Let $\mathbf{P}$ be the transition probability matrix of random walk in a given graph with $\mathbf{P}(i,j) = \mathbf{G}(i,j)/\mathbf{d}_i^{out}$ [30]. The random walk has a stationary distribution $\boldsymbol{\pi}$, which satisfies $\sum_{u_i \in V} \boldsymbol{\pi}(i) = 1$ and $\boldsymbol{\pi}(j) = \sum_{[u_i,u_j]} \boldsymbol{\pi}(i)\mathbf{P}(i,j)$ [30], [31], where $\boldsymbol{\pi}(i) > 0$ for all $u_i \in V$.

To model the social network information, the basic idea is to make the latent representations of two users as close as possible if there exists a following relation between them. It can be mathematically formulated as minimizing

$$\begin{aligned}\mathcal{R}_N = \; & \frac{1}{2} \sum_{[u_i,u_j] \in E} \boldsymbol{\pi}(i)\mathbf{P}(i,j)\|\mathbf{U}_i - \mathbf{U}_j\|^2 \\ = \; & Tr(\mathbf{U}^T(\mathbf{\Pi} - \frac{\mathbf{\Pi}\mathbf{P} + \mathbf{P}^T\mathbf{\Pi}}{2})\mathbf{U}) \\ = \; & Tr(\mathbf{U}^T \triangle \mathbf{U}), \qquad (9)\end{aligned}$$

where $\mathbf{U}_i$ denotes the low-rank representation of user u_i, $\mathbf{U}_j$ the low-rank representation of user u_j, $\triangle = \mathbf{\Pi} - \frac{\mathbf{\Pi}\mathbf{P}+\mathbf{P}^T\mathbf{\Pi}}{2}$ is the Laplacian matrix [31], and $\mathbf{\Pi}$ denotes a diagonal matrix with $\mathbf{\Pi}(i,i) = \boldsymbol{\pi}(i)$. It is straightforward to verify that the Laplacian matrix $\triangle$ has the properties introduced in Lemma (1) and Remark (1). The induction of Eq. (9) is straightforward and can be also found in previous work [31], [30]. This loss function will incur a penalty if two users have different low-rank representations when they have a directed relation.

We have introduced the modeling of content, sentiment and social network information above. Now we propose to consider all of the three types of information in a general framework.

D. Using Sentiment Analysis for Social Spammer Detection

As illustrated in Figure 2, we employ sentiment and network information to formulate two constraints on the matrix factorization model which is derived from content information. By considering all of the three types of information, the task of social spammer detection with sentiment information can be formulated as the following optimization problem:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{V} \geq 0} \quad & \mathcal{O} = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \alpha Tr(\mathbf{U}^T\mathcal{L}\mathbf{U}) \\ & + \beta Tr(\mathbf{U}^T \triangle \mathbf{U}) + \lambda(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2), \qquad (10)\end{aligned}$$

where the first term is to consider content information, the second term is to introduce sentiment information, the third term is to introduce social network information, and the fourth term is for regularization to avoid overfitting. The three positive parameters α, β and λ are to control the effects of each part to the learned model.

The objective function defined in Eq. (10) is not convex with respect to the two variables $\mathbf{U}$ and $\mathbf{V}$ together. There is no closed-form solution for the problem. Motivated by the multiplicative and alternating updating rules discussed in [32], we now introduce an alternative algorithm to find optimal solutions for the two variables $\mathbf{U}$ and $\mathbf{V}$. The key idea is to optimize the objective with respect to one variable, while fixing the other. The algorithm will keep updating the variables until convergence. Now we introduce the algorithm in detail.

1) Computation of $\mathbf{U}$*:* Optimizing the objective function in Eq. (10) with respect to $\mathbf{U}$ is equivalent to solving

$$\begin{aligned}\min_{\mathbf{U}\geq 0} \quad & \mathcal{O}_U = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \alpha Tr(\mathbf{U}^T\mathcal{L}\mathbf{U}) \\ & + \beta Tr(\mathbf{U}^T\triangle\mathbf{U}) + \lambda\|\mathbf{U}\|_F^2,\end{aligned} \tag{11}$$

Let Λ_U be the Lagrange multiplier for constraint $\mathbf{U} \geq 0$, the Lagrange function $L(\mathbf{U})$ is defined as follows:

$$\begin{aligned}L(\mathbf{U}) = & \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \alpha Tr(\mathbf{U}^T\mathcal{L}\mathbf{U}) \\ & + \beta Tr(\mathbf{U}^T\triangle\mathbf{U}) + \lambda\|\mathbf{U}\|_F^2 - Tr(\Lambda_U\mathbf{U}^T),\end{aligned} \tag{12}$$

By setting the derivative $\nabla_{\mathbf{U}}L(\mathbf{U}) = 0$, we get

$$\Lambda_U = -2\mathbf{X}\mathbf{V} + 2\mathbf{U}\mathbf{V}^T\mathbf{V} + 2\alpha\mathcal{L}\mathbf{U} + 2\beta\triangle\mathbf{U} + 2\lambda\mathbf{U}. \tag{13}$$

The Karush-Kuhn-Tucker complementary condition [33] for the nonnegativity constraint of $\mathbf{U}$ gives

$$\Lambda_U(i,j)\mathbf{U}(i,j) = 0 \ ; \tag{14}$$

thus, we obtain

$$[-\mathbf{X}\mathbf{V} + \mathbf{U}\mathbf{V}^T\mathbf{V} + \alpha\mathcal{L}\mathbf{U} + \beta\triangle\mathbf{U} + \lambda\mathbf{U}](i,j)\mathbf{U}(i,j) = 0. \tag{15}$$

Since the Laplacian matrices $\mathcal{L}$ and $\triangle$ may take any signs, we decompose it as $\mathcal{L} = \mathcal{L}^+ - \mathcal{L}^-$ and $\triangle = \triangle^+ - \triangle^-$. Similar to [26], it leads to the updating rule of $\mathbf{U}$,

$$\mathbf{U}(i,j) \leftarrow \mathbf{U}(i,j)\sqrt{\frac{[\mathbf{X}\mathbf{V} + \alpha\mathcal{L}^-\mathbf{U} + \beta\triangle^-\mathbf{U}](i,j)}{[\mathbf{U}\mathbf{V}^T\mathbf{V} + \alpha\mathcal{L}^+\mathbf{U} + \beta\triangle^+\mathbf{U} + \lambda\mathbf{U}](i,j)}}. \tag{16}$$

2) Computation of $\mathbf{V}$*:* Optimizing the objective function in Eq. (10) with respect to $\mathbf{V}$ is equivalent to solving

$$\min_{\mathbf{V}\geq 0} \quad \mathcal{O}_V = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \lambda\|\mathbf{V}\|_F^2, \tag{17}$$

Let Λ_V be the Lagrange multiplier for constraint $\mathbf{V} \geq 0$, the Lagrange function $L(\mathbf{V})$ is defined as follows:

$$L(\mathbf{V}) = \|\mathbf{X} - \mathbf{U}\mathbf{V}^T\|_F^2 + \lambda\|\mathbf{V}\|_F^2 - Tr(\Lambda_V\mathbf{V}^T), \tag{18}$$

By setting the derivative $\nabla_{\mathbf{V}}L(\mathbf{V}) = 0$, we get

$$\Lambda_V = -2\mathbf{X}^T\mathbf{U} + 2\mathbf{V}\mathbf{U}^T\mathbf{U} + 2\lambda\mathbf{V}. \tag{19}$$

The Karush-Kuhn-Tucker complementary condition [33] for the nonnegativity constraint of $\mathbf{U}$ gives

$$\Lambda_V(i,j)\mathbf{V}(i,j) = 0 \ ; \tag{20}$$

thus, we obtain

$$[-\mathbf{X}^T\mathbf{U} + \mathbf{V}\mathbf{U}^T\mathbf{U} + \lambda\mathbf{V}](i,j)\mathbf{V}(i,j) = 0. \tag{21}$$

Similar to [26], it leads to the updating rule of $\mathbf{V}$,

$$\mathbf{V}(i,j) \leftarrow \mathbf{V}(i,j)\sqrt{\frac{[\mathbf{X}^T\mathbf{U}](i,j)}{[\mathbf{V}\mathbf{U}^T\mathbf{U} + \lambda\mathbf{V}](i,j)}}. \tag{22}$$

The correctness and convergence of the updating rules can be proven with the standard auxiliary function approach introduced in [26], [32]. Once obtaining the low-rank user representation $\mathbf{U}$, a supervised model can be trained based on the new latent topic space and label matrix $\mathbf{Y}$. We employ the widely used Least Squares [34], which has a closed-form solution: $\mathbf{W} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Y}$. We present the detailed algorithm of *SDS* in Algorithm 1.

Algorithm 1: *Social Spammer Detection with Sentiment Information*

```
Input: $\{\mathbf{X}, \mathbf{Y}, \mathbf{G}, \alpha, \beta, \lambda, I\}$
Output: U, V, W

1: Construct matrices $\mathcal{L}$ and $\triangle$ in Eq. (7) and (9)
2: Initialize $\mathbf{U}, \mathbf{V} \geq 0$
3: while Not convergent and iter $\leq I$ do
4:     Update
       $\mathbf{U}(i,j) \leftarrow \mathbf{U}(i,j)\sqrt{\frac{[\mathbf{X}\mathbf{V}+\alpha\mathcal{L}^-\mathbf{U}+\beta\triangle^-\mathbf{U}](i,j)}{[\mathbf{U}\mathbf{V}^T\mathbf{V}+\alpha\mathcal{L}^+\mathbf{U}+\beta\triangle^+\mathbf{U}+\lambda\mathbf{U}](i,j)}}$
5:     Update
       $\mathbf{V}(i,j) \leftarrow \mathbf{V}(i,j)\sqrt{\frac{[\mathbf{X}^T\mathbf{U}](i,j)}{[\mathbf{V}\mathbf{U}^T\mathbf{U}+\lambda\mathbf{V}](i,j)}}$
6:     $iter = iter + 1$
7: end while
8: $\mathbf{W} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{Y}$
```

In the algorithm, we conduct initialization for Laplacian matrices, encoding matrix $\mathbf{U}$ and mixing matrix $\mathbf{V}$ from line 1 to 2. I is the number of maximum iterations. The two matrices $\mathbf{U}$ and $\mathbf{V}$ are updated with the updating rules until convergence or reaching the number of maximum iterations. The classifier $\mathbf{W}$ for social spammer detection is trained in line 8.

V. Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of the proposed framework *SDS*. Through the experiments, we aim to answer the following two questions,

1) How effective is the proposed framework compared with other social spammer detection methods?
2) What are the effects of the sentiment information for social spammer detection performance?

We begin by introducing the experimental setup and then compare the performance of different social spammer detection methods. Finally, we study the effects of sentiment information and the parameters on the proposed framework.

A. Experimental Setup

We follow standard experiment settings used in [16], [17] to evaluate the performance of spammer detection methods. We apply different social spammer detection methods on social media datasets. To avoid bias brought by different class distributions, the two Twitter datasets introduced in Section III-A, TUSH and TSS, are used in the experiments. Similar to the literature, precision, recall, and F$_1$-measure are used as the performance metrics.

Three positive parameters are involved in the experiments, including α, β and λ in Eq. (10). α is to control the contribution of sentiment information, β is to control the contribution of social network information, and λ is the regularization parameter to prevent overfitting. As a common practice, all the parameters can be tuned via cross-validation with a separate

TABLE III. SOCIAL SPAMMER DETECTION RESULTS ON TUSH DATASET

	Training Data One (50%)			*Training Data Two (100%)*		
	Precision	Recall	F_1-measure (gain)	Precision	Recall	F_1-measure (gain)
Content_Net	0.893	0.924	0.908 (N.A.)	0.919	0.942	0.930 (N.A.)
Content_Lap	0.926	0.939	0.932 (+2.67%)	0.931	0.949	0.940 (+1.03%)
SMFSR	0.935	0.939	0.937 (+3.12%)	0.948	0.945	0.946 (+1.74%)
SparseSD	0.951	0.955	0.953 (+4.93%)	0.959	0.961	0.960 (+3.17%)
SDS	0.969	0.965	0.967 (+6.47%)	0.975	0.979	0.977 (+5.01%)

TABLE IV. SOCIAL SPAMMER DETECTION RESULTS ON TSS DATASET

	Training Data One (50%)			*Training Data Two (100%)*		
	Precision	Recall	F_1-measure (gain)	Precision	Recall	F_1-measure (gain)
Content_Net	0.801	0.860	0.829 (N.A.)	0.809	0.866	0.837 (N.A.)
Content_Lap	0.821	0.882	0.850 (+2.53%)	0.851	0.902	0.876 (+4.69%)
SMFSR	0.834	0.895	0.863 (+4.10%)	0.860	0.909	0.884 (+5.65%)
SparseSD	0.848	0.900	0.873 (+5.28%)	0.881	0.916	0.898 (+7.37%)
SDS	0.869	0.909	0.889 (+7.12%)	0.898	0.930	0.914 (+9.23%)

validation dataset. In the experiments, we empirically set $\alpha = 0.1$, $\beta = 0.1$ and $\lambda = 0.1$ for general experiment purposes. We empirically set $k = 20$ for k-nearest neighbor defined in Eq. (5). The effects of the parameters on the learning model will be further discussed in Section V-D.

B. Performance Evaluation

We now compare the proposed framework with other baseline methods, accordingly answer the first question asked above. Four baseline methods are included in the experiments:

- *Content_Net*: the content matrix $\mathbf{X}$ and adjacency matrix $\mathbf{G}$ of the social network are combined together for user representation. The basic idea here is to consider each friend of a user as a social dimension [35] for representation. We further use the widely used classifier Least Squares [22] to perform social spammer detection.
- *Content_Lap*: social network information is modeled and incorporated into a Least Squares formulation with a directed Laplacian regularization [30].
- *SMFSR*: a multi-label informed latent semantic indexing [17], [36] is used to model the content information, and undirected graph Laplacian [28] is used to incorporate the social network information. In the experiment, we convert the directed graph to an undirected one with $\mathbf{G} = max(\mathbf{G}, \mathbf{G}^T)$.
- *SparseSD*: a sparse learning framework [37] is used to model the content information, and a directed graph Laplacian [30] is used to incorporate the network information. In the experiment, the directed graph $\mathbf{G}$ is used to model social network information.
- *SDS*: our proposed framework.

Experimental results of the methods on the two Twitter datasets, THSH and TSS, are respectively reported in Table III and IV. In the experiment, we use five-fold cross validation for all the methods. To avoid bias brought by the sizes of the training data, we conduct two sets of experiments with different numbers of training samples. In each round of the cross validation, "Training Data One (50%)" means that we randomly chose 50% of the 80%, thus using 40% of the whole dataset for training. "Training Data One (100%)" represents that we use all the 80% data for training. Also, "gain" represents the percentage improvement of the methods in comparison with the first baseline method *Content_Net*. In the experiment, each result denotes an average of 10 test runs. By comparing the spammer detection performance of different methods, we draw the following observations:

(1) From the results in the tables, we can observe that our proposed method *SDS* consistently outperforms other baseline methods on both datasets with different sizes of training data. Our method achieves better results than the state-of-the-art method *SMFSR* and *SparseSD* on both datasets. We apply two-sample one-tail t-tests to compare *SDS* to the four baseline methods. The experiment results demonstrate that the proposed model performs significantly better (with significance level $\alpha = 0.01$) than the four methods.

(2) The performance of *SDS* is better than the four baselines, which are based on different strategies of utilizing content and network information. This demonstrates that the integration of sentiment information positively helps improve social spammer detection performance.

(3) Among the four baseline methods, *SMFSR* and *SparseSD* achieve better results than the first two methods *Content_Net* and *Content_Lap*. Dimensionality reduction and sparse learning methods show good performance in our studied problem. This indicates that the excellent modeling of content information significantly helps the performance of social spammer detection.

(4) The first method *Content_Net* has the worst performance among all of the four baseline methods. This shows that the proper use of social network information is important in social spammer detection. Simple combination of network information does not work well.

With the help of sentiment information, our proposed framework outperforms the methods incorporating content and network information. Next, we further investigate the effects of sentiment information on the social spammer detection task.

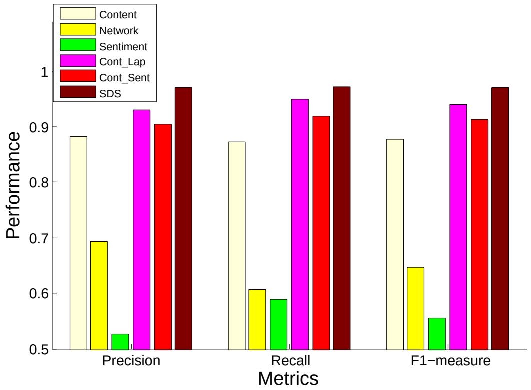

Fig. 3. Spammer Detection Results on TUSH Dataset

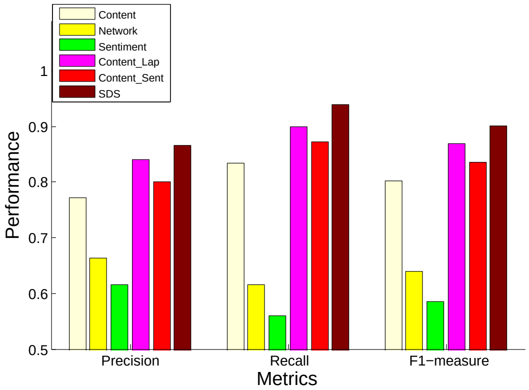

Fig. 4. Spammer Detection Results on TSS Dataset

C. Effects of Sentiment Information

In this subsection, we compare the effectiveness of different types of information to better understand the role of sentiment information in social spammer detection, and accordingly answer the second question asked in the beginning of this section. In particular, we compare the proposed method with the following:

- *Content*: the Least Squares is employed to train a classifier based on only content matrix $\mathbf{X}$.
- *Network*: each friend of a user is considered as a social dimension [35] to represent the user. This is a widely used scheme in relational learning and community detection for user representation. We then train a classifier based on the user-friend representation for social spammer detection.
- *Sentiment*: we first compute the sentiment score of each user and then compare its distance with the mean of spammer group and normal user group. The user is classified into the group with shorter distance.
- *Content_Lap*: the baseline is the same as that in Section V-B.
- *Content_Sentiment*: sentiment information we modeled in Section IV-B is combined with content information for social spammer detection.
- *SDS*: our proposed method to exploit sentiment information for social spammer detection.

The experimental results of the methods on the two datasets are respectively plotted in Figure 3 and 4. In the figures, the first five bars represent the performance of the baselines with different combinations of the information, respectively. The last bar represents our proposed method *SDS*. From the figures, we can draw the following observations:

(1) With the integration of all the three different types of information in a unified way, the proposed framework *SDS* consistently achieves better performance than those with only content and network information. It demonstrates that our proposed method successfully makes use of useful information sources to perform effective social spammer detection.

(2) Among all of the five baseline methods, *Content_Lap* and *Content_Sentiment* achieve better performance than the first three methods. The results indicate that the integration of either network information or sentiment information into a content-based method improves the purely content-based social spammer detection performance. Comparing with traditional spammer detection methods, the use of contextual information positively helps social spammer detection performance.

(3) Among the first three methods, *Content* achieves best performance. This result has been little reported in existing work. It suggests that among the three types of information, content information is the most effective one for social spammer detection. This observation is consistent with those obtained in other platforms, such as email spam detection and Web spam detection. We can observe that *Sentiment* achieves the worst performance, which indicates that we cannot only rely one sentiment information for social spammer detection. Although we observe that the sentiment differences do exist between spammers and normal users, sentiment information is not good enough to be an independent information source to detect spammers.

In summary, the use of sentiment information can help improve the performance of social spammer detection, although it does not work well as an independent information source. The superior performance of the proposed method *SDS* validates its excellent use of the three types of information.

D. Parameter Analysis

As discussed in Section V-A, the effects of two important parameters, i.e., α and β, need to be further explored. α is to control the contribution of sentiment information, and β is to control the contribution of social network information to the model. To better understand the effects brought by the two parameters, we now conduct experiments to compare the social spammer detection performance of the proposed *SDS* on the Twitter datasets with different parameter settings.

The spammer detection results of *SDS* with different parameter settings on the TSS dataset is plotted in Figure 5. From the figure, we can observe that *SDS* achieves relatively good performance when $\alpha < 1$ and $\beta < 1$. When $\alpha > 1$ and $\beta > 1$, as the parameters grow, the performance of *SDS* declines. The results demonstrate that the proposed framework can achieve a

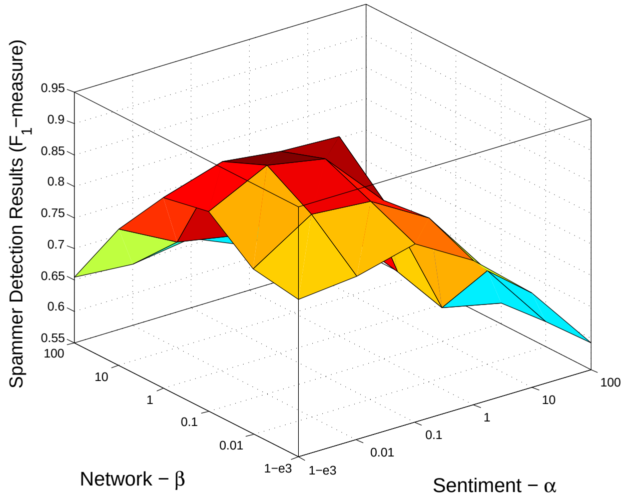

Fig. 5. Impact of Sentiment Information (α) and Social Network Information (β) to the Proposed Framework

relatively good performance when choosing parameter settings in a reasonable range. The performance of *SDS* is not quite sensitive to the parameters. In practice, setting α and β in [0.01, 1] achieves good performance in both datasets. Similar results can be observed on the TUSH dataset; we omit the results owing to lack of space.

VI. Related Work

In this paper, we investigate a novel problem that leverages sentiment information for spammer detection in social media. There are several research areas that are related to our work.

(1) Spammer Detecion. Spammer detection on various platforms, e.g., email [38] and the Web [39], have been studied for years. The spams are designed to corrupt the user experience by spreading ads or driving traffic to particular web sites [39]. A popular and well-developed approach for anti-spam applications is learning-based filtering. The basic idea is that we extract effective features from the labeled data and build a classifier. We then classify new users / messages as either spam or ham according to their content information.

(2) Spammer Detection in Social Media. There are significant efforts to detect and analyze social spammers in Facebook [40], Twitter [41], [42], Renren [16], etc. Following spammer detection in traditional platforms, some work [6] has been done to study tweet content and user behavior for spammer detection in social media. By understanding spammer activities in social networks, features are extracted to perform effective spammer detection. However, the behaviors of the spammers in social media evolve too fast to avoid being detected by a traditional systems that use extensive offline feature building [43].

Another way for social spammer detection is to utilize the social network information [12]. This method is based on the assumption that spammers cannot establish an arbitrarily large number of social trust relations with normal users. This assumption might not hold in many social networks. Yang *et al.* [16] studied the spammers in Renren, the largest OSN in China similar in features to Facebook. Their results reveal that spammers on Renren can have their friend requests accepted by many normal users and thus well blend into the Renren social graph. A similar result targeting Facebook is reported in [40], where the term "social bots" instead of spammers is used. In contrast to Facebook-like OSNs, microblogging systems feature unidirectional user bindings because anyone can follow anyone else without prior consent from the followee. Ghosh *et al.* [41] show 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 following them. Some methods [17] have also proposed to collectively use content and social network information in social spammer detection.

(3) Sentiment Analysis in Social Media. Sentiment analysis on product reviews has been a hot topic for quite a few years [20]. Recently, the opinion-rich resources in social media attracted attention from disciplines. As an effective tool to understand opinions of the public, sentiment analysis is widely applied in various social media applications [44], including poll rating prediction [45], event prediction [46], etc. O'Connor *et al.* [45] found strong correlation between the aggregated sentiment and the manually collected poll ratings. Bollen *et al.* [47] proposed to measure the dynamic sentiments on Twitter, and compared the correlation between public sentiments and major events, including the stock market, crude oil prices, elections and Thanksgiving. Motivated by the successful applications of sentiment analysis and the existing psychological theories, we investigate the use of sentiment information for social spammer detection in this paper.

(4) Opinion Spam Detection. It is popular for people to read opinions for various purposes, such as buying a product or visiting a restaurant. Positive opinions can lead to significant financial gains and/or fames for organizations and individuals. This gives good incentives for opinion spam [48]. Opinion spam detection is an important research topic in sentiment analysis and opinion mining [20]. The objective of this task is to detect spam activities in comments about news articles, blogs, or reviews about products or movies. Our studied problem is different from opinion spam detection. First, we aim to examine spam users in stead of spam review texts, which are often assumed to be independent and identically distributed (i.i.d.). Second, we study a general social spammer detection problem, while opinion spams are always topic-oriented.

VII. Conclusion and Future Work

Social spamming has become a serious problem in almost all kinds of social media services. The distinct characteristics of social media services present new challenges for social spammer detection. Motivated by psychological findings, in this paper, we propose to make use of sentiment information to help social spammer detection. In particular, we first conduct exploratory study on two Twitter datasets to examine the sentiment differences between spammers and normal users. Our experiment results show that the sentiments posed by spammers and normal users are significantly different. The sentiment information are then modeled with a graph Laplacian and incorporated into an optimization formulation. The proposed method considers sentiment, content and network information in a unified way for social spammer detection. Extensive experiments are conducted. The experimental results demonstrate the effectiveness of the proposed framework as well as the roles of different types of information in social spammer detection.

There are many potential future directions based on this work. It is interesting to investigate the contributions of other contextual information, like social activities and linguistic styles, for spammer detection in social media. Also, understanding and analyzing social spammers is also a promising direction. For example, spammers might share similar geographical or temporal patterns. We can thus develop more efficient algorithm to tackle the cold-start problem in spammer detection by utilizing these important patterns.

Acknowledgments

This material is based upon work supported by, or in part by, the U.S. Army Research Office (ARO) under contract/grant number 025071, the National Science Foundation (NSF) under grant number IIS-1217466, and the Office of Naval Research (ONR) under grant number N000141010091.

References

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

[2] 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.

[3] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@ spam: the underground on 140 characters or less," in *Proceedings of CCS*, 2010.

[4] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media," in *ICWSM*, 2011.

[5] 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 *WWW*, 2011.

[6] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," in *Proceedings of SIGIR*, 2010.

[7] J. Weng, E. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of WSDM*, 2010.

[8] E. A. Haggard and K. S. Isaacs, "Micro-momentary facial expressions as indicators of ego mechanisms in psychotherapy." 1966.

[9] P. Ekman, *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. WW Norton & Company, 2009.

[10] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[11] D. Matsumoto, H. S. Hwang, L. Skinner, and M. Frank, "Evaluating truthfulness and detecting deception," *FBI Law Enforcement Bulletin, June*, pp. 1–11, 2011.

[12] P. Boykin and V. Roychowdhury, "Leveraging social networks to fight spam," *Computer*, vol. 38, no. 4, pp. 61–68, 2005.

[13] X. Hu, J. Tang, and H. Liu, "Leveraging knowledge across media for spammer detection in microblogging," in *SIGIR*, 2014.

[14] K. Lee, B. D. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on twitter," in *ICWSM*, 2011.

[15] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of WWW*, 2010.

[16] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Zhao, and Y. Dai, "Uncovering social network sybils in the wild," in *Proceedings of IMC*, 2011.

[17] Y. Zhu, X. Wang, E. Zhong, N. Liu, H. Li, and Q. Yang, "Discovering spammers in social networks," in *AAAI*, 2012.

[18] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Technical Report, Stanford*, 2009.

[19] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of EMNLP*, 2002.

[20] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures on Human Language Technologies*, 2012.

[21] X. Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging," in *WSDM*, 2013.

[22] J. Friedman, T. Hastie, and R. Tibshirani, "The elements of statistical learning," 2008.

[23] X. Hu, N. Sun, C. Zhang, and T.-S. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," in *CIKM*, 2009.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the JMLR*, vol. 3, pp. 993–1022, 2003.

[25] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, pp. 788–791, 1999.

[26] Q. Gu, J. Zhou, and C. H. Ding, "Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs." in *SDM*, 2010, pp. 199–210.

[27] J. Tang, H. Gao, X. Hu, and H. Liu, "Exploiting homophily effect for trust prediction," in *Proceedings of WSDM*, 2013.

[28] F. Chung, *Spectral graph theory*. Amer Mathematical Society, 1997, no. 92.

[29] Y. Lu, M. Castellanos, U. Dayal, and C. Zhai, "Automatic construction of a context-aware sentiment lexicon: an optimization approach," in *Proceedings of WWW*, 2011.

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

[31] F. Chung, "Laplacians and the cheeger inequality for directed graphs," *Annals of Combinatorics*, vol. 9, no. 1, pp. 1–19, 2005.

[32] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *NIPS*, 2001.

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

[34] C. Lawson and R. Hanson, *Solving least squares problems*. SIAM, 1995, vol. 15.

[35] L. Tang and H. Liu, "Relational learning via latent social dimensions," in *Proceedings of KDD*, 2009.

[36] K. Yu, S. Yu, and V. Tresp, "Multi-label informed latent semantic indexing," in *Proceedings of SIGIR*, 2005.

[37] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l 2, 1-norm minimization," in *UAI*, 2009.

[38] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," *Artificial Intelligence Review*, vol. 29, no. 1, pp. 63–92, 2008.

[39] S. Webb, J. Caverlee, and C. Pu, "Introducing the webb spam corpus: Using email spam to identify web spam automatically." in *CEAS*, 2006.

[40] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The socialbot network: when bots socialize for fame and money," in *ACSAC*, 2011, pp. 93–102.

[41] S. Ghosh, B. Viswanath, F. Kooti, N. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. Gummadi, "Understanding and combating link farming in the twitter social network," in *Proceedings of WWW*, 2012.

[42] X. Hu, J. Tang, Y. Zhang, and H. Liu, "Social spammer detection in microblogging," in *IJCAI*, 2013.

[43] X. Hu, J. Tang, and H. Liu, "Online social spammer detection," in *AAAI*, 2014.

[44] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in *WWW*, 2013.

[45] B. O Connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proceedings of ICWSM*, 2010.

[46] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, 2011.

[47] J. Bollen, A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," *CoRR, abs/0911.1583*, 2009.

[48] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of WSDM*, 2008.