Adaptive Spammer Detection with Sparse Group Modeling

Liang Wu[†], Xia Hu[‡], Fred Morstatter[†], Huan Liu[†]
[†]Computer Science and Engineering, Arizona State University, USA
[‡]Department of Computer Science and Engineering, Texas A&M University, USA
{wuliang, fred.morstatter, huanliu}@asu.edu, xiahu@tamu.edu

Abstract

Social spammers disseminate unsolicited information on social media sites that negatively impacts social networking systems. To detect social spammers, traditional methods leverage social network structures to identify the behavioral patterns hidden in their social interactions. They focus on accounts that are affiliated with groups comprising known spammers. However, since different parties are emerging to generate various spammers, they may form different kinds of groups, and some spammers may even detach from the flock. Therefore, it is challenging for existing methods to find the optimal group structure that captures different spammers simultaneously. Employing different approaches for specific spammers is time-consuming, and it also lacks the adaptivity of dealing with emerging spammers.

In this work, we aim to propose a group modeling framework that adaptively characterizes social interactions of spammers. In particular, we introduce to integrate content information into the group modeling process. The proposed framework exploits additional content information in selecting groups and individuals that are likely to be involved in spamming activities. In order to alleviate the intensive computational cost, we transform the problem as a sparse learning task that can be solved efficiently. Experimental results on real-world datasets show that the proposed method outperforms the state-of-the-art approaches.

1 Introduction

Social media sites have become a popular platform for information dissemination. The increasing availability and popularity of social media sites, combined with the potential for automation, allow for the rapid creation and spread of spam, which unfairly overwhelms legitimate users with unwanted information. Since spamming behaviors significantly hinder the overall value of social systems going forward, detection of spammers would positively influence the quality of social networking services and user experience.

Existing efforts have been made to discover the behavioral patterns of social spammers deviating from that of legitimate users. In addition to messages, social networking services make it available to utilize network structures to detect spammers. Traditional methods can be classified into three categories, *i.e.,* link-based, neighbor-based, and group-based methods. Link-based methods utilize links as a measure of social trust, where links are assumed to be established by legitimate users. Therefore, simple measures like the number of followers and followee/follower ratio can directly be used to detect spammers. Neighbor-based methods utilize links as a measure of homophily, assuming that links are established within legitimate users and spammers. However, these methods face novel challenges due to evolved spamming strategies.

Since many users follow back when they are followed by someone for the sake of courtesy, spammers could establish a decent number of links with legitimate users (Sedhai and Sun 2015). These noisy links no longer represent social trust or a sign of homophily, which undermine the performance of link- and neighbor-based methods. A more robust way is to model the network structure instead of individual links, which has been studied for identifying product review spammers (Ye and Akoglu 2015). However, social network structures are more difficult to be modeled, since social spammers generated by different parties may detach or form a group, and the group granularity could also be various.

In this work, we present a framework that adaptively identifies different spammers by integrating content and network structures. In particular, we hierarchically represent social network users with the social group structures. Observing that a large group could usually be split into several subsystems, we find many splits are unnecessary for spammer detection. Therefore, we pose the spammer detection problem as a sparse learning task and leverage the additional content information to search the optimal group structures catering to spammer detection. The proposed method, Sparse Group Modeling for Adaptive Spammer Detection (*SGASD*), adaptively detects not only different spammer groups but also individual spammers detached from the spammer flocks. The main contributions of this paper are outlined as follows:

- Introduce an emerging problem of spammer detection, which cannot be solved by existing solutions;
- Present a novel framework to adaptively model different types of social interactions of spammers;
- Suggest mathematical formulation to solve the optimization problem efficiently; and

- Evaluate the proposed method on real world Twitter datasets and elaborate the effects of different components.

The remainder of this paper is structured as follows. In Section 2, we introduce the problem of social spammer detection formally. In Section 3, we introduce our proposed framework, as well as the optimization method and theoretical analysis. We conduct experiments on real world datasets in Section 4. We introduce related work in Section 5. In Section 6, we conclude the paper and present future work.

2 Problem Statement

Let $\mathbf{A} = [\mathbf{V}, \mathbf{P}, \mathbf{t}]$ be a target account set with social media content $\mathbf{V}$, social links $\mathbf{P}$ and labels $\mathbf{t}$. Specifically, $\mathbf{V} \in \mathbb{R}^{m \times n}$ are the *n* content features of *m* users; $\mathbf{P} \in \mathbb{R}^{m \times m}$ denotes the adjacency matrix, where $\mathbf{P}_{i,j} = 1$ indicates that the j^{th} account is followed by the i^{th} account and equals to 0 otherwise. $\mathbf{t} \in \{0, 1\}^m$ records the identity label of a subset of users, where $\mathbf{t}_i = 0$ represents the i^{th} account is a spammer and equals 1 otherwise.

Given a set of social media actors $\mathbf{A}$ *with their attribute matrix* $\mathbf{V}$*, social links* $\mathbf{P}$*, identity label information* $\mathbf{t}$ *of part of users in the dataset, our goal is to learn a model* $\mathbf{w} \in \mathbb{R}^n$ *with best performance to classify whether an unknown account i is a spammer or not with* $\mathbf{V}_{i,*}\mathbf{w}$.

3 Adaptive Spammer Detection

In this section, we will first introduce how we enable the adaptive group structure modeling to be efficient, and then introduce how content information could be integrated. Next, we will introduce the corresponding optimization algorithm of *SGASD*. Finally, we provide a theoretical analysis of time complexity as well as its convergence.

3.1 Adaptive Group Structure Modeling

Finding group structure of linked nodes has been extensively studied in the literature of community detection, which is usually reduced to a clustering problem. The clustering process is usually time-consuming. An adaptive group structure is challenging to obtain since the update of parameters might require an entire re-computation. We propose to solve this problem by first generating all possible groups, and then select optimal combinations.

In order to generate more possible groups, an intuitive idea is to incrementally cluster all users hierarchically (Fortunato 2010). Namely, individual users are first gathered to form small groups, and small groups are merged to bigger ones. Therefore, these (final and intermediate) groups provide necessary group structures with different granularities. In order to represent the social group index, we introduce the concept of index tree T, which is defined as follows,

Definition 1. ***Index tree*** T*: Let* T *denote a tree of depth d, where non-leaf nodes represent social communities and leaf nodes are users. Let* $T_i = \{G_1^i, G_2^i, \ldots, G_{n_i}^i\}$ *denote the nodes on layer i, where* $n_0 = 1$ *and* n_i *is the number of nodes on layer i. Given* $i < d$*,* G_j^i *represents* j^{th} *group on the* i^{th} *layer.* $G_1^0 = \{1, 2, \ldots, m\}$ *contains indices of all users. In order to maintain a tree structure, nodes should satisfy the following conditions: 1) Nodes on the same layer share no indices with each other* ($G_j^i \cap G_k^i = \emptyset, \forall i = 1, \ldots, d, j \neq k, j \leq n_i, k \leq n_i$)*; 2) Given a non-root node* G_j^i*, we denote its parent node as* G_{j0}^{i-1} ($G_j^i \subseteq G_{j0}^{i-1}, 1 < i \leq d$).

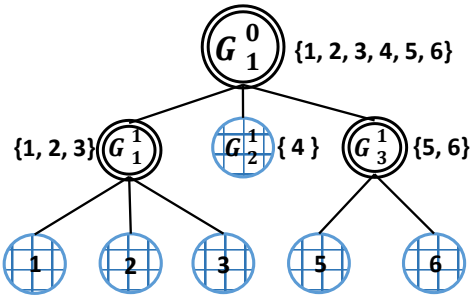


Figure 1: An illustrative example of social community structures, where the hierarchical communities are of various resolutions and the leaf node represents the individual user.

Fig. 1 illustrates a toy example of community structures with various resolutions, where the nodes filled with blue lines are leaf nodes. In order to obtain such a group structure, we select a hierarchical community detection method, namely Louvain (Blondel et al. 2008), where maximum modularity is used to optimize the group structure. The code is publicly available[1].

Next, we introduce how we select groups with sparse representation. Sparse learning aims to achieve sparse representation of data, allowing only discriminant elements to be selected, while noisy and redundant ones are discarded. In order to employ sparse learning, we introduce a weighting vector $\mathbf{c} \in \mathbb{R}^m$, where each entry is the weight for a user. The sparse representation of the group structure can then be obtained through minimizing

$$\min_{\mathbf{c}} \sum_{i=0}^{d} \sum_{j=1}^{n_i} ||\mathbf{c}_{G_j^i}||_2, \tag{1}$$

where an ℓ_2-$norm$ is imposed on each member of a group, and an ℓ_1-$norm$ is imposed on weights of all groups. The combination of ℓ_1- and ℓ_2-$norm$ leads to sparse representation of $\mathbf{c}$, while ℓ_1-$norm$ determines the organization of sparsity. In particular, imposing ℓ_1-$norm$ within each group leads to the inter-group sparsity, *i.e.,* weights of users in some groups are selected to be assigned higher weights, while users in other groups are with lower weights. Therefore, redundant and noisy groups are filtered by the sparse representation of $\mathbf{c}$.

3.2 Spammer Detection with Group Structures

Since content information is also useful for identifying spammers. In this section, we introduce how we integrate social media content into the group modeling framework.

[1] https://perso.uclouvain.be/vincent.blondel/research/louvain.html

For generality, we adopt a regression method to model messages as follows,

$$\frac{1}{2}\min_{\mathbf{w}}||\mathbf{V}\mathbf{w}-\mathbf{y}||_2^2+\frac{\lambda_1}{2}||\mathbf{w}||_2^2, \tag{2}$$

where $\mathbf{V} \in \mathbb{R}^{m\times n}$ is the data matrix, m is the number of users and n is the number of textual features. $\mathbf{V}_{i,j}$ represents the j^{th} feature of the i^{th} user. The feature value can be induced based on term frequency using different measures, such as normalized term frequency and TF-IDF. In this work, only the term frequency is used. $\mathbf{w} \in \mathbb{R}^n$ is the model which needs to be optimized. $\mathbf{y} \in \mathbb{R}^m$ is the label vector of training data. The formulation achieves an optimal $\mathbf{w}$ through minimizing the training error. $||\mathbf{w}||_2^2$ is the regularizer avoiding overfitting. λ_1 controls the extent of simplicity of the model. Note that other supervised methods can also be used here.

Next, we employ the information induced from group memberships to regularize the predictor as follows,

$$\begin{aligned}&\min_{\mathbf{w},\mathbf{c}}\frac{1}{2}\sum_{i=1}^{m}\mathbf{c}_i(\mathbf{V}_{i,*}\mathbf{w}-\mathbf{y}_i)^2+\frac{\lambda_1}{2}||\mathbf{w}||_2^2\\&\text{subject to}\quad\sum_i\mathbf{c}_i=1,\end{aligned} \tag{3}$$

where each user is weighted by $\mathbf{c}_i$. We aim to achieve an optimal $\mathbf{c}$ which differentiates selected groups of legitimate users and spammers. By incorporating the regularizer in Eq.(1), the objective can then be reformulated as follows,

$$\begin{aligned}&\min_{\mathbf{w},\mathbf{c}}\frac{1}{2}\sum_{i=1}^{m}\mathbf{c}_i(\mathbf{V}_{i,*}\mathbf{w}-\mathbf{y}_i)^2+\frac{\lambda_1}{2}||\mathbf{w}||_2^2+\lambda_2\sum_{i=0}^{d}\sum_{j=1}^{n_i}||\mathbf{c}_{G_j^i}||_2\\&\text{subject to}\quad\sum_i\mathbf{c}_i=1,\end{aligned} \tag{4}$$

where λ_2 controls extent of inter-group sparseness, meaning that a larger λ_2 leads to fewer groups being selected. Next, we will introduce how parameters can be learned efficiently.

3.3 Learning Parameters

There are two variables that need to be optimized in Eq.(4). Since the objective is not jointly convex with respect to both $\mathbf{c}$ and $\mathbf{w}$, we propose to find optimal solutions through alternatively updating them, *i.e.*, fixing one and updating the other. By reducing the problem into two convex optimization tasks, $\mathbf{c}$ and $\mathbf{w}$ keep being updated until convergence. Now we introduce details of the algorithm:

Learning the predictor When $\mathbf{c}$ is fixed, the problem only depends on $\mathbf{w}$. We reformulate the objective function as follows:

$$\epsilon_{\mathbf{w}}=\frac{1}{2}\sum_{i=1}^{m}\mathbf{c}_i(\mathbf{V}_{i,*}\mathbf{w}-\mathbf{y}_i)^2+\frac{\lambda_1}{2}||\mathbf{w}||_2^2. \tag{5}$$

Therefore, the problem is reduced to an ℓ_2 regularized weighted linear regression problem, which is to minimize the cost $\epsilon_{\mathbf{w}}$. Since social media users and their corresponding contents may be massive, a scalable optimization method is needed. Here we use Stochastic Gradient Descent (SGD) (Bottou 2010). Since Eq.(5) is convex, the corresponding gradient can directly be obtained as:

$$\frac{\partial\epsilon_{\mathbf{w}}}{\partial\mathbf{w}}=\sum_{i=1}^{m}\mathbf{c}_i\mathbf{V}_{i,*}^T(\mathbf{V}_{i,*}\mathbf{w}-\mathbf{y}_i)+\lambda_1\mathbf{w}. \tag{6}$$

SGD is scalable since data examples can be updated in parallel (Zinkevich et al. 2010). Detailed discussions about the performance can be found in Section 4.

Learning the group structure When $\mathbf{w}$ is fixed, Eq.(4) depends only on $\mathbf{c}$. Since the squared loss $(\mathbf{V}\mathbf{w}-\mathbf{y}_i)^2$ becomes a constant, we replace it with $\mathbf{p}$, where $\mathbf{p}_i=(\mathbf{V}_{i,*}\mathbf{w}-\mathbf{y}_i)^2$. The objective can then be reformulated as:

$$\begin{aligned}&\min_{\mathbf{c}}\frac{1}{2}\sum_{i=1}^{m}\mathbf{c}_i\mathbf{p}_i+\lambda_2\sum_{i=0}^{d}\sum_{j=1}^{n_i}||\mathbf{c}_{G_j^i}||_2\\&\text{subject to}\quad\sum_i\mathbf{c}_i=1,\end{aligned} \tag{7}$$

where the regularizer $||\mathbf{w}||_2^2$ that is fixed here is also omitted. It is easy to prove that Eq.(7) is strongly convex but not directly differentiable, *i.e.*, it is convex and non-smooth with respect to $\mathbf{c}$. In order to find the solution for the optimization problem in Eq.(7), we reformulate the problem as follows:

$$\phi_{\lambda_2}(\mathbf{c})=\arg\min_{\mathbf{c}}\frac{1}{2}||\mathbf{c}-\mathbf{x}||^2+\lambda_2\sum_{i=0}^{d}\sum_{j=1}^{n_i}||\mathbf{c}_{G_j^i}||_2, \tag{8}$$

where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{x}_i=\frac{\mathbf{p}_i^{-1}}{\sum_k^m\mathbf{p}_k^{-1}}$. Therefore, the equality constrained optimization problem is transformed to a Moreau-Yosida regularization problem with the euclidean projection of $\mathbf{c}$ on to a vector $\mathbf{x}$ (Lemaréchal and Sagastizábal 1997). The new formulation is continuously differentiable and it admits an analytical solution (Liu and Ye 2010). Given a proper λ_2, the optimal $\mathbf{c} \in \mathbb{R}^m$ can be obtained in an agglomerative manner, which is shown in Algorithm 1. In the algorithm, the superscript of $\mathbf{c}$ is used to denote the layer of the tree, meaning that the output of the algorithm is $\mathbf{c}^0$. The bisection method can be implemented to find the optimal λ_2. Empirically, λ_2 can be initialized as $\sqrt{\frac{||l'(\mathbf{0})||_2^2}{\sum_{i=0}^{d}n_i}}$, where $l(c)=\frac{1}{2}||\mathbf{c}-\mathbf{x}||^2$. Then we use $\phi_{\lambda_2}(-l'(\mathbf{0}))$ to test whether λ_2 is large or small enough. When $\phi_{\lambda_2}(-l'(\mathbf{0}))=\mathbf{0}$, which means λ_2 is large enough to generate a trivial solution, we start looking for the lower bound as follows:

$$\lambda_2^{(lower)}=\max\{\lambda_2^{(i)}|\lambda_2^{(i)}=\frac{\lambda_2^{(i)}}{2^i},\pi_{\lambda_2^{(i)}}(-l'(\mathbf{0}))\neq\mathbf{0}\} \tag{9}$$

otherwise, if $\phi_{\lambda_2}(-l'(\mathbf{0}))\neq\mathbf{0}$, we start looking for the upper bound as follows:

$$\lambda_2^{(upper)}=\min\{\lambda_2^{(i)}|\lambda_2^{(i)}=2^i\lambda_2^{(i)},\pi_{\lambda_2^{(i)}}(-l'(\mathbf{0}))=\mathbf{0}\} \tag{10}$$

Algorithm 1: Solution of Moreau-Yosida Regularization

Input: $\{\mathbf{c}, G, \lambda_2\}$
Output: $\mathbf{c}^0$.
1: Set $\mathbf{c}^{d+1} = \mathbf{x}$,
2: **for** $i = d$ to 0 **do**:
3: **for** $j = 1$ to n_i **do**:
4: Compute:

$$\mathbf{c}^i_{G^i_j} = \begin{cases} \mathbf{0} & \text{if} \quad ||\mathbf{c}^{i+1}_{G^i_j}||_2 \leq \lambda_2, \\ \frac{||\mathbf{c}^{i+1}_{G^i_j}||_2 - \lambda_2}{||\mathbf{c}^{i+1}_{G^i_j}||} \mathbf{c}^{i+1}_{G^i_j} & \text{if} \quad ||\mathbf{c}^{i+1}_{G^i_j}||_2 > \lambda_2, \end{cases}$$

5: **end for**
6:**end for**

In Algorithm 1, we traverse the tree in an agglomerative manner, *i.e.*, from leaf nodes to the root node. At each node, the ℓ_2-$norm$ of the weight $\mathbf{c}$ can be reduced by at most λ_2 as shown in step 4. After the traverse, the analytical solution of $\mathbf{c}$ can be achieved.

3.4 Time Complexity Analysis

Here we analyze the time complexity of the algorithm. The computational costs include computation of $\mathbf{c}$ and $\mathbf{w}$. The computational cost for $\mathbf{c}$ comes from estimating the Moreau-Yosida regularization problem, which takes $\sum_{i=0}^{d}\sum_{j=1}^{n_i}|G^i_j|$. The computation of $\mathbf{w}$ is a standard ℓ_2 regularized regression problem, which can be accelerated with the parallel implementation. The calculation of Louvain method could also speed up and it needs to be done only once as preprocessing. Since the optimization is conducted in an alternative manner and both sub-tasks are convex, both procedures will monotonically decrease. In addition, since the objective function has lower bounds, such as zero, the above iteration converges.

4 Experiments

In this section, experiments are conducted to test the effectiveness of *SGASD*. In particular, we compare *SGASD* with state-of-the-art approaches based on real-world datasets.

4.1 Datasets

In the literature of spammer detection, two methods are commonly used to obtain data with ground truth, *i.e.*, using Twitter suspended user list, and using social honeypot accounts. We adopt two Twitter datasets by each of them.

The first dataset (TwitterS) is collected using the suspended account list of Twitter. In order to get ground truth data for positive instances, we follow the conventional practice to use the suspended user list. We crawled data from February 3^{rd}, 2011 to February 21^{st}, 2013, and follow the conventional practice (Hu et al. 2013) to check whether they were suspended as spammers. The second dataset (TwitterH) is obtained from a subset of followers of honeypot accounts (Lee, Eoff, and Caverlee 2011) and it is publicly available[2]. Statistics of the datasets are shown in TABLE 1 and Table 2.

Table 1: Statistics of TwitterS dataset used in this study. Labels are obtained through Twitter suspended users.

Tweets	Unique Users	ReTweets
1,150,192	94,535	576,167
Links	**Spammers**	**Positive Ratio**
23,047,758	7,061	7.5%

Table 2: Statistics of TwitterH dataset used in this study. Labels are obtained through followers of honeypot accounts.

Tweets	Unique Users	ReTweets
4,453,380	38,400	223,115
Links	**Spammers**	**Positive Ratio**
8,739,105	19,200	50%

4.2 Baseline Methods

In this section, we adopt several methods to evaluate different aspects of *SGASD*. We include state-of-the-art link-based, neighbor-based and group-based methods for spammer detection, as well as two variations of *SGASD* to evaluate different components. Details of the adopted baselines are as follows:

Ratio of Follower/Followee: The ratio of follower/followee (RFF) is a metric that identifies anomalous social behaviors in the link-based manner (Lee, Eoff, and Caverlee 2011).

Social Spammer Detection in Microblogging: Social Spammer Detection in Microblogging (*SSDM*) (Hu et al. 2013) incorporates a graph regularizer into Elastic Net, which proves to be effective in dealing with spammers in a neighbor-based manner.

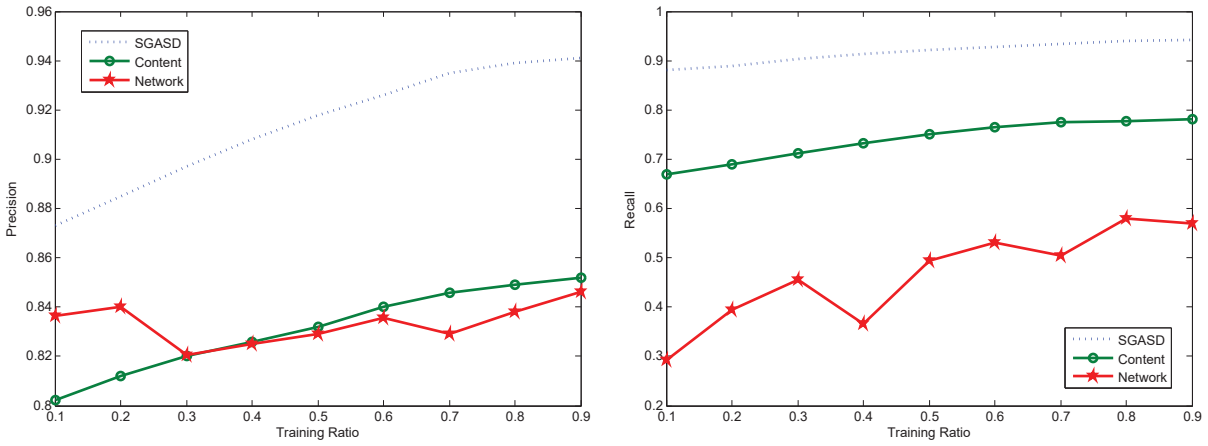
Network Footprints Score: The approach detects groups of spammers by measuring the Network Footprints Score (*NFS*) (Ye and Akoglu 2015), which is a metric quantifying the distortion of reviews brought by spammers towards a product. We adopt this method by replacing product rating scores with the concentration probability of topics, since social spammers aim to lead attention to a particular topic.

Content: We propose a variant of *SGASD* that only models the message by removing the tree-structured regularizer from Eq.(4). The method is a least square regression model.

Network: We propose another variant of *SGASD* that only models the network structure. In particular, we use Louvain method to find groups of both training and testing examples. The testing examples which are 1) detached from all groups; 2) members of spammer-dominated groups; are classified as spammers.

As a common practice, parameters of these methods, such as the decision probability threshold for RFF, are tuned through cross-validation on separate validation datasets.

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



(a) Precision with varying ratio of spammers.

(b) Recall with varying ratio of spammers.

Figure 2: Precision and recall of *SGASD*, *Content* and *Network* with varying ratio of positive and negative examples.

Table 3: The precision, recall and $\mathbf{F_1}$-measure comparison on the TwitterS dataset. *SGASD* achieves the best $\mathbf{F_1}$.

Method	Precision	Recall	F_1
RFF	51.75%	56.80%	54.16%
$SSDM$	87.66%	80.26%	83.80%
NFS	84.31%	66.23%	74.18%
$Content$	80.21%	66.93%	72.97%
$Network$	83.63%	29.35%	43.45%
$SGASD$	87.30%	88.72%	**88.01%**

Table 4: The precision, recall and $\mathbf{F_1}$-measure comparison on TwitterH dataset, where *SGASD* achieves the best $\mathbf{F_1}$.

Method	Precision	Recall	F_1
RFF	77.60%	65.02%	70.76%
$SSDM$	92.15%	92.00%	92.07%
NFS	88.16%	65.67%	75.27%
$Content$	86.80%	82.46%	84.58%
$Network$	76.30%	45.10%	56.68%
$SGASD$	93.75%	96.92%	**95.31%**

4.3 Experimental Results

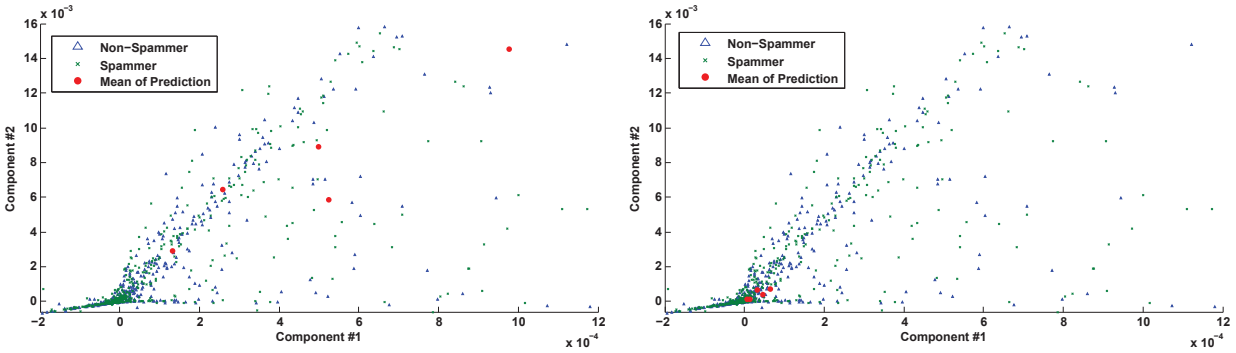
Experimental results on TwitterS and TwitterH can be found in TABLE 3 and TABLE 4, and all results below are obtained through 10-fold cross-validation, where the reported result is the average of the ten folds.

Based on the results, we make following observations. Link-based approaches RFF cannot achieve appealing results with real-world data since spammers may gain enough links even with regular users. $SSDM$ delivers the runner-up performance, so it would fit some particular spammers with certain patterns, but is not generalizable for a wider range of spammers. NFS outperforms RFF by learning links jointly from the group perspective. $SGASD$ achieves the best result by adaptively modeling the group structures. $SGASD$ also outperforms the two variants $Network$ and $Content$, by jointly instead of separately learning them.

Since a fundamental difference between TwitterS and TwitterH is the different extent of data skewness, it would be interesting to investigate the sensitivity of *SGASD* to varying ratios of positive and negative examples. In the second experiment, we use TwitterS data and vary the ratio of spammers against non-spammers through randomly down-sampling non-spammers. The ratio ranges from 10% to 90%. Furthermore, we also test *Content* and *Network* to understand how different components of *SGASD* work toward varying skewness. Experimental results are shown in Figure 2.

Based on the results, we make following observations. As shown in Figure 2(a), precision of *Content* increases steadily with more spammers. Since it only considers content information in messages, when there are more positive instances and fewer negative instances being put, it is easier to identify patterns of spammers regarding content. As shown in Figure 2(b), recall of *Content* increases slightly with more spammers in the data, while that of *Network* increases rapidly. Though such increase of *Network* is unstable, *SGASD* can take advantage of both information and achieves the optimal recall and improves rapidly.

In order to further investigate the difference brought by the adaptive group modeling, we plot the prediction results of *SGASD* and the runner-up method *SSDM* for comparison in Figure 3. In particular, we rank testing data examples according to their prediction scores, and then group top 250 examples by every 50 examples and plot the mean of each group using the red circles. We also plot the testing points with different labels. Here, the x- and the y-axis is the first and second component of the data correspondingly, which are obtained through Principle Component Analysis. It can be seen that *SSDM* keeps focusing on the dense area where spammers are dominant, while *SGASD* also detects spammers in the sparse area. It shows that *SGASD* can



(a) Centers of results with highest confidence of *SGASD*. (b) Centers of results with highest confidence of *SSDM*.

Figure 3: An example of difference on prediction results between *SGASD* and *SSDM*.

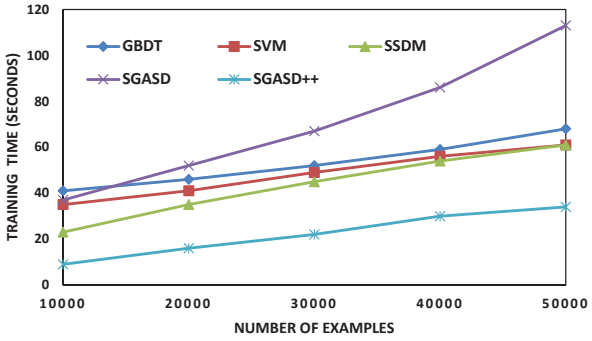


Figure 4: Training time of methods on varying size of training data with the same convergence condition. *SGASD++* learns parameters with 8 threads in parallel.

detect spammers with less apparent patterns.

4.4 Scalability Studies

In order to investigate the scalability of the method, we report the performance of spammer detection methods and two conventional classification algorithms with varying size of data in Figure 4. *SGASD++* is optimized with 8 threads. As shown in the figure, multi-threaded optimization of $\mathbf{w}$ significantly accelerates the training, which achieves a sub-linear time cost with increasing data instances. The experiment is performed on a 3.40 GHz Dual-Core Intel Core i7 CPU.

SGASD can be viewed as imposing the adaptive group analysis on *Content* information. In order to further investigate the efficiency of group analysis, we also report the time cost of Louvain method and *Content* in TABLE 5. It can be seen that the running time of *SGASD* adds around 10 to 15 extra seconds to *Content*. Although Louvain method takes more time, its computation could be conveniently accelerated (Low et al. 2014) and needs to be done only once as preprocessing.

In summary, the proposed approach *SGASD* can identify spammers more effectively (F_1) than all baseline methods through optimizing both precision and recall. Network and content information contribute to different aspects of spammer detection. Through integrating content analysis with adaptive group modeling, *SGASD* can utilize the network structure better and detect spammers both effectively and efficiently.

Table 5: Running time (*seconds*) of group analysis methods and the content-based method.

Methods \ # Posts	1×10^5	2×10^5	3×10^5	5×10^5
Louvain	56	93	122	178
Content	33	46	61	98
SGASD	37	52	67	113

5 Related Work

The network modeling methods can generally be divided into three categories, link-based, neighbor-based and group-based, which are illustrated in Fig. 5. Link-based methods assume links are generally regarded as social trust from other users, and a small number of links might indicate a spammer being fake. The underlying assumption is that social media users are carefully connected, which might not be true in the real world. Since users would simply follow back after being followed, social media users with more followees are found to own more followers generally. A revised solution is to compile features such as the ratio of follower/followee (Lee, Eoff, and Caverlee 2011). However, spammers could follow users incrementally and unfollow those who did not follow back seeminglessly, which is transient and difficult to notice.

"Birds of a feather flock together", neighbor-based methods regard links as a sign of homophily, assuming that linked users are more likely to share the same label. The homophily constraint could be incorporated in two ways. First, links could be directly used to clamp prediction results of connected users (Rayana and Akoglu 2015). Second, since spammers usually focus on specific topics, a small set of features are assumed to be more effective in discovering spam. Links can then be used to smooth selection of features by preserving the similarity between connected users in the

diminished feature space (Hu et al. 2013; Li et al. 2016). However, sophisticated spammers could defeat related filters through link farming.

Group-based methods leverage the group structure hidden in social networks to detect spammers' social interactions, which are generally more robust to noisy links. Prior work on incorporating social groups can generally be divided into two categories. First, since attacks of social spammer are coordinated to launch, they are assumed to form a spammer group *w.r.t.* their links and messages (Jindal and Liu 2007). Second, the problem of spammer detection could also be reduced to that of outlier detection (Akoglu, Tong, and Koutra 2015), since spammers are 'outliers' in the sense of behaving differently from non-spammers (Gao et al. 2010). Note that the first category aims to achieve a group structure where spammers are clustered together, while the second aims to achieve a group structure where spammers are detached to any clusters. Each of the two categories only focuses the specific spammers. Moreover, parameters of groups such as group size and number of groups are crucial to the performance, which are difficult to optimize over a large information network with massive features and links. *SGASD* jointly considers spammers with both social interaction patterns and can adaptively model the group structure.

The content-based methods aim to discover patterns from posts of malicious users (Wu et al. 2016). The patterns can be drawn from friend requests, private messages and comments (Jindal and Liu 2007; Morstatter et al. 2016; Wu et al. 2017a). As social media platforms provide users with a variety of activities to participate in, various features have also been used in previous spammer detection research. Correlations between similar posts and correlations between historical and present data have been utilized to find misinformation (Sampson et al. 2016; Wu et al. 2017b) Since accounts of spammers are often generated in a batch, similar or even same profile templates are often used. Webb *et al.* tried to discover these accounts through learning the user profiles (Webb, Caverlee, and Pu 2008). As spammers are often employed to post information related to a specific topic, their posting behavior often contains long hibernation and bursty peaks. Chu *et al.* proposed related approach to leverage the temporal feature and discover spammers (Chu et al. 2010). User behavioral patterns, such as online rating (Lim et al. 2010), temporal burstiness (Ye, Kumar, and Akoglu 2016), locations (Li et al. 2015), friend invitation (Xue et al. 2013) and social ties (Yang et al. 2014) have also been studied.

Our work is also related to sparse learning methods. Sparse learning was proposed for generating sparse representations of models and data instances, where ℓ_1-norm is often adopted to regularize the parameters. In real applications, sparsity with some specific structure is often found to be effective. In order to encode different kinds of structures, a variety of sparse learning methods have been proposed, such as group lasso and fused lasso. The structured sparsity is achieved through using ℓ_1-norm on different levels. For example, Kong *et al.* proposed to encode intra-group sparsity of models through introducing $\ell_{1,2}$-$norm$ (Kong et al. 2014), where the $\ell_{1,2}$-$norm$ is imposed on instances inside each group (Wu, Hu, and Liu 2016). Thus, a small set of features are selected in each group. Liu *et al.* proposed to impose sparsity toward tree-structured groups and select certain groups of features (Liu and Ye 2010), which is different from the regularizer in our work for selecting groups of instances. In addition to features, our method also clusters data examples, which distances itself from multiple task grouping methods (Daumé III 2013)

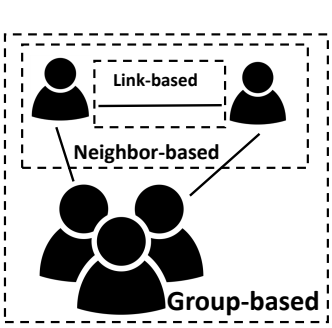


Figure 5: An illustrative example of how the social network structure has been used in traditional spammer detection.

6 Conclusion and Future Work

Social media platforms have been used by spammers to overwhelm normal users with unwanted information. Various methods have been proposed to discover the specific patterns of spammers and detect them, however, it becomes more challenging since sophisticated spammers find ways to establish links with legitimate users, and thus are able to trick existing social spammer algorithms. In this work, we reviewed how network information has been used in existing spammer detection methods, and we found that adaptive methods might be useful in dealing with these noisy links. In order to allow for the adaptive detection of spammers, we propose a novel sparse group modeling method to characterize group structures hidden in social networks. Through leveraging the redundancy of social groups, and content information of training data, the proposed framework *SGASD* can detect spammers effectively and efficiently. Empirical results are obtained based on two real world Twitter datasets.

Several possible future directions remain to be studied. Since we focus on automatically generated labeled datasets, it would be interesting to test the performance of different models on datasets which are labeled by human annotators, such as through crowdsourcing platforms. Social media platforms enable users to share various kinds of information, such as texts, user names, friends, time and locations. We mainly focus on leveraging texts and links between users, it would be interesting to design a framework which is able to infer identity by incorporating all these heterogeneous information sources.

Acknowledgments

The work is funded, in part, by ONR N00014-16-1-2257 and the Department of Defense under the MINERVA initiative through the ONR N000141310835.

References

Akoglu, L.; Tong, H.; and Koutra, D. 2015. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery* 29(3):626–688.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10):P10008.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer. 177–186.

Chu, Z.; Gianvecchio, S.; Wang, H.; and Jajodia, S. 2010. Who is tweeting on twitter: human, bot, or cyborg? In *26th Annual Computer Security Applications Conference*, 21–30. ACM.

Daumé III, A. K. H. 2013. Learning task grouping and overlap in multi-task learning.

Fortunato, S. 2010. Community detection in graphs. *Physics Reports* 486(3):75–174.

Gao, J.; Liang, F.; Fan, W.; Wang, C.; Sun, Y.; and Han, J. 2010. On community outliers and their efficient detection in information networks. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 813–822. ACM.

Hu, X.; Tang, J.; Zhang, Y.; and Liu, H. 2013. Social spammer detection in microblogging. In *International Joint Conference on Artificial Intelligence*, 2633–2639.

Jindal, N., and Liu, B. 2007. review spam detection . In International Conference on World Wide Web, 1189–1190.

Kong, D.; Fujimaki, R.; Liu, J.; Nie, F.; and Ding, C. 2014. Exclusive feature learning on arbitrary structures via l12-norm. In *NIPS*, 1655–1663.

Lee, K.; Eoff, B. D.; and Caverlee, J. 2011. Seven months with the devils: A long-term study of content polluters on twitter. In *AAAI Conference on Web and Social Media*.

Lemaréchal, C., and Sagastizábal, C. 1997. Practical aspects of the moreau–yosida regularization: Theoretical preliminaries. *Journal on Optimization* 7(2):367–385.

Li, H.; Chen, Z.; Mukherjee, A.; Liu, B.; and Shao, J. 2015. Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In *AAAI Conference on Web and Social Media*.

Li, J.; Hu, X.; Wu, L.; and Liu, H. 2016. Robust unsupervised feature selection on networked data. In *SIAM International Conference on Data Mining*, 387–395. SIAM.

Lim, E.-P.; Nguyen, V.-A.; Jindal, N.; Liu, B.; and Lauw, H. W. 2010. Detecting product review spammers using rating behaviors. In *International Conference on Information and Knowledge Management*, 939–948. ACM.

Liu, J., and Ye, J. 2010. Moreau-yosida regularization for grouped tree structure learning. In *NIPS*, 1459–1467.

Low, Y.; Gonzalez, J. E.; Kyrola, A.; Bickson, D.; Guestrin, C. E.; and Hellerstein, J. 2014. Graphlab: A new framework for parallel machine learning. *arXiv preprint arXiv:1408.2041*.

Morstatter, F.; Wu, L.; Nazer, T. H.; Carley, K. M.; and Liu, H. 2016. A new approach to bot detection: striking the balance between precision and recall. In *ASONAM*, 533–540. IEEE ACM.

Rayana, S., and Akoglu, L. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, 985–994. ACM.

Sampson, J.; Morstatter, F.; Wu, L.; and Liu, H. 2016. Leveraging the implicit structure within social media for emergent rumor detection. In *International on Conference on Information and Knowledge Management*, 2377–2382. ACM.

Sedhai, S., and Sun, A. 2015. Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. In *SIGIR Conference on Research and Development in Information Retrieval*, 223–232. ACM.

Webb, S.; Caverlee, J.; and Pu, C. 2008. *Social Honeypots: Making Friends With A Spammer Near You*. Conference on Email and Anti-Spam.

Wu, L.; Morstatter, F.; Hu, X.; and Liu, H. 2016. Chapter 5: Mining misinformation in social media. In *Big Data in Complex and Social Networks*. CRC Press. 123–152.

Wu, L.; Hu, X.; Morstatter, F.; and Liu, H. 2017a. Detecting camouflaged content polluters. In *International AAAI Conference on Web and Social Media*.

Wu, L.; Li, J.; Hu, X.; and Liu, H. 2017b. Gleaning wisdom from the past: Early detection of emerging rumors in social media. In *SIAM International Conference on Data Mining*. SIAM.

Wu, L.; Hu, X.; and Liu, H. 2016. Relational learning with social status analysis. In *International Conference on Web Search and Data Mining*, 513–522. ACM.

Xue, J.; Yang, Z.; Yang, X.; Wang, X.; Chen, L.; and Dai, Y. 2013. Votetrust: Leveraging friend invitation graph to defend against social network sybils. In *INFOCOM*, 2400–2408. IEEE.

Yang, Z.; Wilson, C.; Wang, X.; Gao, T.; Zhao, B. Y.; and Dai, Y. 2014. Uncovering social network sybils in the wild. *TKDD* 8(1):2.

Ye, J., and Akoglu, L. 2015. Discovering opinion spammer groups by network footprints. In *Machine Learning and Knowledge Discovery in Databases*. Springer. 267–282.

Ye, J.; Kumar, S.; and Akoglu, L. 2016. Temporal opinion spam detection by multivariate indicative signals. *arXiv preprint arXiv:1603.01929*.

Zinkevich, M.; Weimer, M.; Li, L.; and Smola, A. J. 2010. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, 2595–2603.