

Community-Based Features for Identifying Spammers in Online Social Networks

Sajid Yousuf Bhat*

*Department of Computer Science
Jamia Millia Islamia, New Delhi, India
Email: s.yousuf.bhat@gmail.com

Muhammad Abulaish*^{†‡}, SMIEEE

[†]Center of Excellence in Information Assurance
King Saud University, Riyadh, Saudi Arabia
Email: abulaish@ieee.org

Abstract—The popularity of Online Social Networks (OSNs) is often faced with challenges of dealing with undesirable users and their malicious activities in the social networks. The most common form of malicious activity over OSNs is spamming wherein a bot (fake user) disseminates content, malware/viruses, etc. to the legitimate users of the social networks. The common motives behind such activity include phishing, scams, viral marketing and so on which the recipients do not intend to receive. It is thus a highly desirable task to devise techniques and methods for identifying spammers (spamming accounts) in OSNs. With an aim of exploiting social network characteristics of community formation by legitimate users, this paper presents a community-based framework to identify spammers in OSNs. The framework uses community-based features of OSN users to learn classification models for identification of spamming accounts. The preliminary experiments on a real-world dataset with simulated spammers reveal that proposed approach is promising and that using community-based node features of OSN users can improve the performance of classifying spammers and legitimate users.

Index Terms—Social network analysis; Social network security; Spammer Detection; Community-Based feature identification.

I. INTRODUCTION

Today's emerging technology and trends have resulted in numerous systems and platforms through which social entities interact and communicate with each other; for example, e-mail, Text Messaging, Telephone/Mobile networks and most significantly Online Social Networks (OSNs) like Facebook and Google+. These systems enable their users to join, and share ideas, information, and interests through various forms of interactions supported by them. Additionally, OSNs allow users to publish their personal information (profile), additional multimedia content, and link to other users whom they relate to. The communication and interaction services provided by these systems enable to reveal the underlying social networks of their users and thus they represent a unique opportunity to study and understand them. An in-depth analysis of social network structure and growth can lead to a better design of future social network based systems. Online social networks offer many useful properties that reflect real-world social network characteristics, which include small-world behavior, significant local clustering [1], existence of large strongly connected component [2] and formation of tightly knit groups or communities [3].

The wide popularity of OSNs and their ease of access has also resulted in the misuse of their services. Besides the issue of preserving user privacy, OSNs face the challenge of dealing with undesirable users and their malicious activities in the social network. The most common form of malicious activity identified in OSNs is spamming which involves malicious users (spammers) to broadcast irrelevant information in the form of e-mails, IMs, comments, Text Messages and posts to as large number of legitimate users as possible. Spamming is done mostly with an aim of promoting products, viral marketing, spreading fads, and in some cases may possibly be done to harass legitimate users of an OSN in order to decrease their trust in the particular service. Some of the spamming related issues that are of major concern include:

- Spam interactions utilize large amounts of network bandwidth leading to less revenue and significant financial losses to organizations.
- Spamming leads to uncontrolled dissemination of information, content, malware/viruses, promotional ads, phishing and scams which the recipients do not intend to receive. This could lead to OSN users becoming prey of tricky scams or harassment and lead to their dissatisfaction with the service.

In this regard, it is highly desirable to devise techniques and methods for identifying spammers and their behavior in online social networks. Once identified, spammers can be blocked or removed from the social networks and future spam activity can be significantly controlled by analyzing spammer behavior.

Many spam/spammer detection methods have been proposed in literature, which are based on the content analysis (keywords-based filtering) of the interactions between users. However, many counter-filtering techniques based on the usage of non-dictionary words and images in spam objects are often employed by spammers. Content-based spam filtering systems also demand higher computations. Moreover, the issue of privacy-preservance of user content (private messages, posts, profile details) is often held against content-based spam filtering systems. Alternatively, some spammer detection techniques are based on learning classification models from network-based topological features of the interacting nodes in online social networks. These features mainly include in-degree, out-degree, reciprocity, clustering coefficient, etc.

[‡]To whom correspondence should be addressed. E-mail: abulaish@ieee.org

Spammers are often seen to mimic some patterns of legitimate interaction behavior making it difficult to characterize them. Incorporating additional sociological characteristics like, interaction behavior of nodes within and across network community structures, in the classification models can make it more difficult for spammers to qualify as legitimate nodes and thus improve classification. In this regard, the aim of this paper is to improve spammer classification models by incorporating some community-based features of nodes besides the basic topological features. In this paper, the community structure from interaction graphs of social networks is identified using the density-based overlapping community detection method `OCTracker` proposed in [4]. Various node features are then extracted to learn a classification model from a set of pre-labeled training instances. Moreover, in order to restrict spamming, higher-level communities are identified on a super-graph of the node-level communities. These higher-level communities represent the boundaries within which interactions are considered to be legitimate whereas interactions across higher-level communities can be considered as spam. Results related to the performance of a spammer classification model involving the use of various topological and community-based features are also presented.

II. BACKGROUND AND RELATED WORK

This section presents the motivation and background related to the detection of spammers in OSNs. The main aim here is to present the basic concepts on which the proposed spammer detection approach is based. Moreover, this section also presents a brief review of the recent techniques developed to tackle the detection of spammers.

A. Community Structures

One of the important properties of social networks including OSNs that has been studied with high interest is the clustering property of nodes (users), i.e. the formation of user communities. In a community, the nodes are relatively densely connected to each other but sparsely connected to other dense groups in the network. Identifying community structure in social networks is important as it reveals the functional groups in a system and thus provides information about the role of individual nodes. For example, a node at the boundary of a community may work as an important mediator between communities, whereas a central node may provide control and stability to the community. Traditional community detection techniques like [5] aim to identify distinct/disjoint communities from social networks using various approaches like graph partitioning, hierarchical clustering, modularity optimization and so on. However, individuals in real-world social networks are often seen to participate or belong in multiple overlapping communities. In this regard, a popular method for identifying overlapping communities is the Clique Percolation Method (CPM) proposed by Palla et al. [6] which is based on the concept of a k -clique, i.e., a complete subgraph of k nodes. Other methods dealing with the nature of overlapping communities include [7].

One of the important properties of the real-world social networks is that they tend to change dynamically as most often: i) new users join the network, ii) old users leave the network, and iii) users establish/break ties with other users. Consequently, all these evolutionary events result in birth, growth, contraction, merge, split, and death of communities with time. Although, recent literature includes some approaches for analyzing communities and their temporal evolution in dynamic networks, a common weakness in these studies is that communities and their evolutions have been studied separately. As pointed out in [8], a more appropriate approach would be to analyze communities and their evolution in a unified framework, where community structure provides evidence about community evolution. In this regard, Bhat and Abulaish [4] propose a density-based overlapping community detection method `OCTracker` which tracks the community evolution in dynamic networks by adapting a known community structure (previously identified) to the new topological changes occurring in the network with time.

B. Spammer Properties

On the web the most common form of spamming is the search-engine spamming or spamdexing. It is the form of topological spamming where link farms (densely connected set of pages) are created explicitly with the intention of deceiving a link based ranking algorithm [9]. The basic assumption to deal with spamdexing is that similar objects are related to similar objects in the webgraph. Linked hosts tend to belong to the same class, i.e., either they all are spam or all are non-spam [10]. In the context of OSNs, this type of spamming along with copy-profiling could be done to promote fake influential nodes which may affect the correctness of recommender systems in OSNs.

Considering the case of OSNs the most common form of spamming is the Random Link Attack (RLA) where a small number of spammers send spam to a large number of randomly selected victim nodes. Spammers tend to be senders of spam messages to a socially un-related set of receivers [11], unlike legitimate senders whose receivers tend to cluster or form communities as discussed earlier. It is unlikely that the recipients of the spam messages sent by a spammer have friend or friend-of-friend relations or have some kind of mutual ties among them [12], [13]. As a result, a distinctive feature that has often been used to detect spammers is the clustering coefficient (CC) by considering that networks representing connections of legitimate users show high CC while spammers show CC close to 0 [14]. However, in many cases spammers make their neighborhood structurally similar to legitimate nodes and thus increase their CC, making it hard to detect them [11].

Another detection scheme that is commonly used to stop spam and identify spammers based on collaborative filtering involves using a user voting scheme to classify a message as spam or non-spam. The message recipients are provided with options by their messaging service providers to vote a received message as spam or non-spam. These votes are then

collectively used to identify spamming IP addresses and user accounts [15], [16]. However, a deceptive scheme used by spammers to get away from collaborative filtering spam detection methods is the vote-gaming attack [15], [16] wherein, spammers add some of the secondary accounts controlled by them to the recipients list of spam messages sent from a spamming account. When a secondary account receives a spam message that is already classified as spam, the bot controlling the secondary account will report the message as non-spam. Considering non-spam votes from multiple spammer accounts, the spam filtering system will notice the lack of consensus and not filter the message as spam for other recipients.

One of the unique distinguishing properties between spammers and normal users in OSNs is that the interactions of spammers are least often reciprocated while as, mostly, all of the legitimate user interactions are reciprocated [17], [13]. It may also be the case that a group of spammers fakes communication reciprocity between them by reciprocating each other's interactions which they also send as spam to a comparably small set of legitimate targets so as to increase their reciprocated interaction average. However, in order to be effective as spammers and meet their goals, they need to target as larger number of legitimate nodes as possible. Spamming a small number of legitimate nodes in the system will have a negligible effect on the system. It means that faking interaction reciprocity alone is not a good solution for spammers to deceive a filtering system which considers the interaction reciprocity for detecting spammers.

C. Related Work

Most of the techniques and methods developed for spammer or spam detection from online social networks involve a content based approach. Such approaches learn classification models using various machine learning techniques from known spam instances (training set) based on the textual features of spammer profile details (about me, address and so on) or their interactions (e-mails, messages, wall posts and so on) or both like [18]. The main idea is based around the observation that spammers use distinguished keywords, URLs and so on in their interactions and to define their profiles. However, it is not always true and such an assumption is often deceived by the approaches like copy-profiling and content obfuscation. In order to improve spam/spammer detection, besides textual-features, additional features based on images, topological properties of interaction networks and social network properties have recently been used. For example, DeBarr and Wechsler [19] uses both content and social network metrics like degree centrality based features to learn a classifier for the task. Wang [20] uses graph based metrics to improve spam classification on a microblogging platform. Jin et al. [21] use a combined feature set incorporating heterogeneous features based on images, text and social network behavior of a user profile on an online social network to learn a classification model. Benevenuto et al. [22] aim to identify spammers in video sharing online social networks by incorporating three

sets of attributes, for machine learning, including video attributes (ratings provided to an uploaded video by other users), user attributes (activity on the site) and social network metrics (clustering coefficient, betweenness, reciprocity and so on). Other methods which incorporate a mix of content based and topological features include [23], [10], [17]. Lee et al. [24] define social honeypots (administered bot accounts) that monitor spammers' behaviors and log their information. If the social honeypot detects suspicious user activity (e.g., the honeypot's profile receives a friend request, message, wall post and so on) then the social honeypot's bot collects evidence of the spam candidate. They further use machine learning techniques to learn classification models from the information collected by the social honeypots. However, one of the main limitations of social honeypots is their reach, i.e., not all spammers would target them, and that the classifiers can possibly be deceived if the spammers involve a copy-profile attack (i.e., imitate the profile of a legitimate user). As mentioned earlier, the issues related to user-privacy and computational requirements of content based filtering systems often hints on using only link based, topological and social network properties of the communication networks for identifying spammers. In this regard, Shrivastava et al. [11] incorporate only structural properties which include clustering coefficient and neighborhood independence to deal with the Random Link Attacks from Spammers. Gan and Suel [25] extract only link based features like in-links, out-links, cross-links etc. from a web graph to classify pages as spam or not. Other methods include finding physical node clusters based on network-level features from online communication networks for example, [26]. The methods proposed in [27] and [15] aim to identify vote gaming attacks by considering the voting behavior of users and the IP addresses they use. They follow a graph based clustering approach to identify malicious groups trying to imitate the legitimate behavior. To detect spam clusters, Gao et al. [28] use two widely acknowledged distinguishing features of spam campaigns: their "distributed" coverage and "bursty" nature. The "distributed" property is quantified using the number of users that send wall posts in the cluster. The "bursty" property is based on the intuition that most spam campaigns involve coordinated action by many accounts within short periods of time [29]. Lam [13] shows how communication reciprocity, communication interaction average and clustering coefficient of the nodes in OSNs can be used to differentiate spammers from legitimate users. A motivation for our proposed approach comes from [30] which aims to learn communication patterns (based on reciprocity, clustering coefficient and so on) from the dynamic user interactions and form relation pattern graphs that characterize the behavior of legitimate senders and spammers. Our approach considers dynamic overlapping communities as the pattern graphs and exploits the role of nodes within communities and the interaction behavior of nodes across communities to classify them. Another very closely related work is that of [31] wherein they use a community detection method to split the interaction network into communities. They extract features based on the degree of a user, the number of

communities the user is connected to, number of links between the friends of the user, and the average number of friends inside each of the user’s connected communities. To the best of our knowledge, [31] and the method proposed in this paper are the first steps towards using various community-based features for identifying spammers in OSNs.

III. PROPOSED METHOD

The main aim of the proposed framework is to detect spammers from online social networks. It is based on learning a classification model from community-based features of the nodes after identifying their node level community structure from the weighted interaction graph of the social network. The weight of a directed link in the graph represents the total number of messages, posts, etc., sent from the origin to the destination. The basic idea of the proposed invention is shown in figure 1. The various steps involved in the proposed

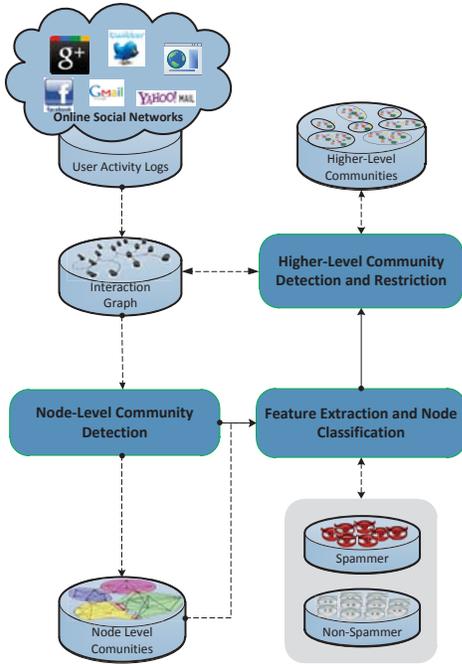


Fig. 1: Overview of the framework

framework are discussed in the following sub-sections.

A. Node Level Community Detection

The framework starts with detecting density-based node level overlapping communities from the interaction graph of online social network users using the OCTracker algorithm [4]. The interaction graph is usually generated from the activity logs of the users like the email logs, wall-post logs and so on. Unlike traditional density-based community detection methods, OCTracker is an overlapping community detection method which also tracks the evolution of communities with time. Moreover, it requires only one input parameter, η , (instead of two as in traditional methods) which can be tuned to identify communities at different resolutions. Some of the

important features of this method that we use include, a) categorizing nodes as cores (important nodes in a community), non-cores (boundary nodes of a community), and outliers (nodes which do not belong to a community), and b) Overlapping nature of nodes, i.e., the number of communities a node assigned to. For details on the community detection method, see the actual paper for the same.

B. Feature Extraction

Once the overlapping community structure of nodes is identified the next step involves extracting community-based and some topological features of nodes in the network. They include the features which express the role of a node in the community structure, i.e., whether a node is a boundary node or a core node and the number of communities it belongs to (if any). It also uses out-degree and reciprocity related node features, however, in the light of community membership. The various features and their description used in this paper are given as follows:

Total out-degree: The total out-degree of a node represents the total number of distinct users in the social network to which it has out links, i.e., sends messages etc.

Total reciprocity: The total reciprocity of a node represents the ratio of the number of nodes with which it has both in-links and out-links, to the total number of nodes to which it has out-links. Formally, for a node p if I_p is the set of nodes which have out-links to p and O_p is the set of nodes to which node p has out-links, then the total reciprocity of p , i.e., TR_p is given by equation 1.

$$TR_p = \frac{|I_p \cap O_p|}{|O_p|} \quad (1)$$

Total in/out ratio: For a node p it represents the ratio of the number of nodes which have out-links to p , i.e., $|I_p|$, to the number of nodes and to which node p has out-links, i.e., $|O_p|$ as given in equation 2.

$$TIOR_p = \frac{|I_p|}{|O_p|} \quad (2)$$

Core node: This is a boolean property which is true for a node p if the community detection method used here, OCTracker, marks the node p as a core-node, otherwise it is false.

Community memberships: This feature represents the number of communities to which the overlapping community detection method, OCTracker, assigns a particular node p . For the outlier nodes, the value for this feature will be zero.

In order to define the following node features, we first define a *foreign node* for any particular node p in our context. For a node p , a node q is called a *foreign node* if the two nodes p and q do not belong to a common community. We now define the other community based node features.

Foreign out-degree: The total number of foreign nodes to which a node p has out-links is called the foreign out-degree of node p represented as $|FO_p|$.

Foreign in/out ratio: The foreign in/out ratio for a node p is defined as the ratio of the number of foreign nodes that have

out-links to the node p , i.e., $|FI_p|$, to the number of foreign nodes to which node p has out-links, i.e., $|FO_p|$ as given in equation 3.

$$FIOR_p = \frac{|FI_p|}{|FO_p|} \quad (3)$$

Foreign out-link probability: This feature represents the probability that a particular node p has an out-link to a foreign node. If FI_p is the set of foreign nodes to which a node p has out-links and O_p is the set of all nodes to which p has out links, then the foreign out-link probability of node p is given by equation 4.

$$FOP_p = \frac{|FI_p|}{|O_p|} \quad (4)$$

Foreign reciprocity: For a node p if FI_p is the set of foreign nodes which have out-links to p and FO_p is the set of nodes to which node p has out-links, then the foreign reciprocity of p , i.e., FR_p is given by equation 5.

$$FR_p = \frac{|FI_p \cap FO_p|}{|FO_p|} \quad (5)$$

Foreign out-link grouping: This feature basically represents the probability that the foreign nodes to which a node p has out-links, i.e., FO_p , have a common community. If $MFO_p \subseteq FO_p$ is the maximal set of nodes that have a common community, then this feature value is calculated as the ratio of the number of nodes in MFO_p to the total number of nodes in FO_p as given in equation 6.

$$FOG_p = \frac{|MFO_p|}{|FO_p|} \quad (6)$$

C. Classification

After extracting the node features the task is to learn a classifier using a set of pre-labeled nodes in the interaction graph that have already been classified as spam or non-spam. These pre-labeled nodes can be the result of administrative spam filtering performed on the basis of either content filtering of profiles and messages, or user reports and feedback about the senders in online social networks. In either case, the community-based features of these pre-labeled nodes form the training set for learning the classifier. In literature, many machine learning methods have been used to learn classifiers based on topological and content-based features of spam and spammers in online social networks. The most commonly used classifiers include NaiveBayes, decision tree and k -NN to name a few. An illustration of the process of learning the classification model from the extracted features is presented in figure 2

The classification model selected from the learning phase is then used to classify un-labeled nodes of the interaction graph representing the online social network. The newly identified spammer nodes are reported to the system administrator who further decides whether to block the suspected nodes or completely remove them from the social network.

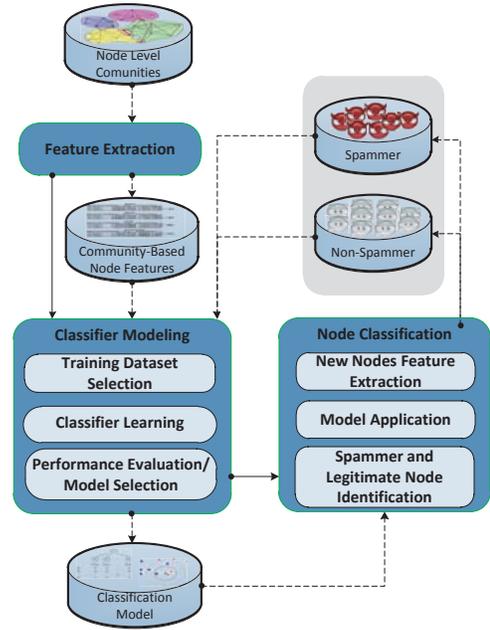


Fig. 2: Classification model learning and application

D. Dealing with future interactions

In order to prevent future spamming in the non-spammer social network we follow a scheme that involves identifying maximal node groups within which nodes interact socially. As mentioned earlier, the users of online social networks interact within a group of other users or small worlds. An interaction outside these groups can be considered as suspicious. However, using only the node level communities of non-spammers may seem to be too restrictive as online social networks show dynamic behavior and evolve with time. As a result we extract higher level communities from the node-level community structure of the interaction graph. The process is illustrated via a block diagram in figure 3.

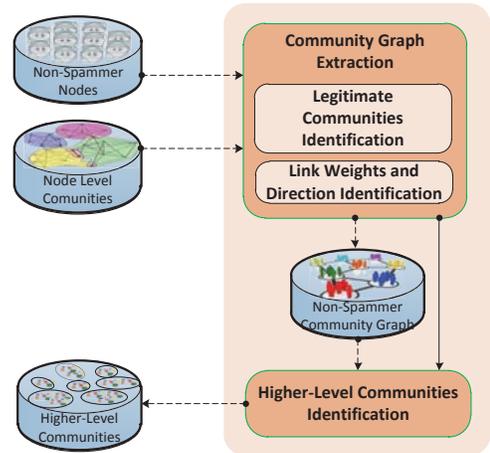


Fig. 3: High-level community structure identification

Considering each node-level community as a single node,

we extract a super-graph by determining the inter-community interaction counts and directions resulting in a directed-weighted community graph. We re-apply the community detection algorithm on this super-graph to find the higher-level community structure of the OSN users. Each higher-level community represents a small world consisting of node-level communities that interact within that small world and are least likely to communicate outside their small world in the near future. According to this setting, any future interaction across these small worlds is considered as suspicious and the sender as a possible spammer as illustrated in figure 4.

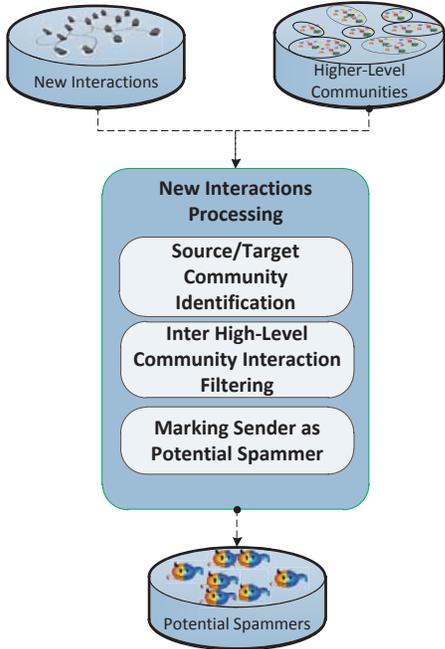


Fig. 4: Future spam prevention by inter high-level community filtering

IV. EXPERIMENTS

The main component of this paper is learning a classifier from the community-based node features of the users of online social networks. In this regard here we present the performance of some classification models learnt using the proposed features. We present the performance of multiple classifiers including decision trees, NaiveBayes and k -NN implemented in the WEKA [32] software on a set of real-world social networks with artificially planted spammer nodes.

A. Datasets

We use two real-world datasets, one representing the wall-post activity of about 63891 Facebook users [33] and another representing the email network of about 87273 users of Enron [34]. We aim to represent the nodes in these two networks as legitimate nodes and inject additional nodes in the networks simulating spammer behavior. In this regard we subsequently filter out all the nodes having zero in-degree or out-degree, and any isolated nodes from the two networks to represent

them as legitimate networks. This results in the Facebook network containing 32693 legitimate nodes and the Enron network containing 9272 legitimate nodes. The reason for such less number of nodes in case of Enron dataset is that the actual network is generated from the inbox of only 150 Enron employees and is thus partial for most of the nodes. Now in order to simulate spammers, we generate a set of 1000 isolated nodes for each legitimate network, which create out-links to randomly selected nodes in the respective legitimate networks. The out-links or the out-degree generated for the spammers are not random but follow the distribution shown by spammers as reported in [12] and also used in [35], [13] as shown in Table I. As earlier mentioned, the messages of the spammers are

TABLE I: Spammer out-degree distribution

y	P[out-degree=y]
1	0.664
2	0.171
3	0.07
4	0.04
5	0.024
6	0.014
7	0.01
8	0.007

expected to be least often reciprocated. Thus the probability of a legitimate node replying to a spammer is set to 0.05. In order to make the detection task more difficult, we generate another set of 1000 spammer nodes, for each legitimate network, which try to mimic the clustering/community property of legitimate nodes. In order to do so, we use the LFR-benchmark generator [36] to generate a directed network of 1000 nodes with embedded community structures. The various LFR-benchmark parameters used to generate the network are shown in Table II.

Now for each node in the synthetic network, we rewire a set of its out-links towards a set of randomly selected nodes in a legitimate network such that the spamming out-degree (i.e., the rewired out-links) follows the distribution given in Table I. In this regard, a total of 2000 spammer nodes (out of which 1000 mimic the clustering property of legitimate nodes) are added to each legitimate network resulting in a total of 34693 nodes for the Facebook network and 11272 nodes for the Enron network. We now apply the overlapping community detection method OCTracker on each dataset and extract the

TABLE II: LFR-Benchmark parameter description and values for spammer network generation

Parameter	Description	Value
N	number of nodes	1000
k	average degree	15
k_{max}	max degree	60
C_{min}	minimum community size	15
C_{max}	maximum community size	60
τ_1	degree exponent	-1
τ_2	community exponent	-1
μ	mixing parameter	0.1

various features for each node in the respective networks.

B. Results

In order to evaluate the significance of our approach, we learn a set of classifiers from WEKA on the training examples containing the community-based features from the datasets mentioned in the previous section. We evaluate the performance of four classifiers including two decision-tree based (J48 [37], ADTree [38]), one k -NN based (IBk [39], using $k=5$ nearest neighbors) and the NaiveBayes [40]. We use 10-fold cross validation for each classifier on the two datasets to evaluate the performance. Table III presents the performance of the various classifiers on the Facebook dataset with planted spammers and Table IV presents their performance on the Enron dataset with the planted spammers. As can be seen from Tables III and IV, the decision-tree based classifiers J48 and ADTree perform better than the others and have a low false-positive rate on both the classes.

TABLE III: Performance on Facebook network with simulated spammers

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
J48	0.973	0.086	0.981	0.973	0.977	0.974	Non-spam
	0.914	0.027	0.881	0.914	0.897	0.974	spam
ADTree	0.947	0.063	0.986	0.947	0.966	0.985	Non-spam
	0.937	0.053	0.791	0.937	0.858	0.985	spam
IBk	0.959	0.171	0.963	0.959	0.961	0.975	Non-spam
	0.83	0.041	0.814	0.83	0.822	0.975	spam
NaiveBayes	0.667	0.117	0.964	0.667	0.788	0.865	Non-spam
	0.883	0.333	0.364	0.883	0.515	0.866	spam

TABLE IV: Performance on Enron network with simulated spammers

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
J48	0.999	0.023	0.999	0.999	0.999	0.996	Non-spam
	0.978	0.001	0.977	0.978	0.977	0.996	spam
ADTree	0.998	0.045	0.997	0.998	0.998	0.999	Non-spam
	0.955	0.002	0.973	0.955	0.964	0.999	spam
IBk	0.998	0.054	0.997	0.998	0.997	0.998	Non-spam
	0.946	0.002	0.964	0.946	0.955	0.998	Spam
NaveBayes	0.958	0.175	0.989	0.958	0.973	0.958	Non-spam
	0.826	0.042	0.543	0.826	0.655	0.959	spam

We also evaluate the case where we use the Facebook dataset as the training set and the Enron dataset as the test set. This is to ensure that a classifier does not show overspecialization on a particular dataset. Table V presents the results for this case using the best classifier (J48) from the previous experiments (i.e., Table III and IV). As can be seen from Table V, the performance is a little degraded but still good enough indicating that the feature set used here can be used to classify spammers and non-spammers in online social networks. In order to further ensure the significance of the various community-based features, used in this paper, for identifying spammers in online social networks we evaluate the performance of J48 classifier using only the non-community based node features, i.e., out-degree, total reciprocity, and total in/out ratio. We generate three results

TABLE V: Performance of J48 classifier on the Facebook network (with simulated spammers) as training set and the Enron network (with simulated spammers) as test set

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
J48	0.894	0.71	0.983	0.894	0.936	0.922	Non-spam
	0.929	0.106	0.654	0.929	0.768	0.922	Spam

as shown in Table VI wherein the first two rows correspond to the performance of J48 classifier on the Facebook dataset using 10-Fold cross validation. The second and the third rows correspond to the performance of J48 classifier on the Enron dataset using 10-Fold cross validation. The last two rows correspond to the performance of J48 classifier using Facebook dataset as the training set and the Enron dataset as the test set. On comparing the results presented in Table VI with the results of the J48 classifier in Tables III, IV and V we can see that using community based features of nodes in online social networks along with the non-community based features in classification shows better performance than simply using the non-community based features.

TABLE VI: Performance of the J48 classifier using only the non-community based features

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
Facebook	0.994	0.074	0.995	0.994	0.995	0.993	Non-spam
	0.926	0.006	0.91	0.926	0.918	0.993	Spam
Enron	0.971	0.172	0.963	0.971	0.967	0.981	Non-spam
	0.828	0.029	0.862	0.828	0.844	0.981	Spam
Facebook (train) Enron (test)	0.817	0.08	0.979	0.817	0.891	0.935	Non-spam
	0.921	0.183	0.52	0.921	0.665	0.935	Spam

V. CONCLUSION AND FUTURE WORK

This paper has presented a community-based framework to identify spammers in online social networks (OSNs). Starting with the identification of overlapping community structures from user interaction network, a set of community-based features are identified to build a classification model for detecting spam nodes in OSNs. The community structures identified from the interaction network are further used to restrict spamming by filtering interactions across higher-level communities identified on a super-graph of node-level communities (wherein each node represents a node-level community). A node (user account), say x , can be labeled as a potential spammer, if it aims to send a message to another nodes that are not within the same higher-level community in which x belongs. Though the proposed approach needs extended evaluation, the preliminary experimental results indicate that the proposed approach is significant. In future, we aim to provide a more extensive evaluation of the complete framework on real-world application datasets.

ACKNOWLEDGMENT

The authors would like to thank King Abdulaziz City for Science and Technology (KACST) and King Saud University for their support. This work has been financially supported by KACST under the NPST project number 11-INF1594-02.

REFERENCES

- [1] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.
- [2] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 611–617.
- [3] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.
- [4] S. Y. Bhat and M. Abulaish, "Octracker: A density-based framework for tracking the evolution of overlapping communities in osns," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Los Alamitos, CA, USA: IEEE Computer Society, 2012, pp. 501–505.
- [5] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, 2004.
- [6] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.
- [7] S. Gregory, "An algorithm to find overlapping community structure in networks," in *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2007, pp. 91–102.
- [8] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng, "Analyzing communities and their evolutions in dynamic social networks," *ACM Trans. Knowl. Discov. Data*, vol. 3, pp. 8:1–8:31, April 2009.
- [9] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates, "Link-based characterization and detection of web spam," in *Second International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Seattle, USA, 2006.
- [10] F. J. Ortega, C. Macdonald, J. A. Troyano, and F. Cruz, "Spam detection with a content-based random-walk algorithm," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, ser. SMUC '10. New York, NY, USA: ACM, 2010, pp. 45–52.
- [11] N. Shrivastava, A. Majumder, and R. Rastogi, "Mining (social) network graphs to detect random link attacks," in *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, ser. ICDE '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 486–495.
- [12] L. H. Gomes, R. B. Almeida, L. M. A. Bettencourt, V. Almeida, and J. M. A., "Comparative graph theoretical characterization of networks of spam and legitimate email," in *Proceedings of the 2nd Conference on Email and Anti-Spam (CEAS)*, 2005.
- [13] H. Lam, *A Learning Approach to Spam Detection Based on Social Networks*. Hong Kong University of Science and Technology, 2007.
- [14] P. O. Boykin and V. P. Roychowdhury, "Leveraging social networks to fight spam," *Computer*, vol. 38, no. 4, pp. 61–68, Apr. 2005.
- [15] A. Ramachandran, A. Dasgupta, N. Feamster, and K. Weinberger, "Spam or ham?: characterizing and detecting fraudulent "not spam" reports in web mail systems," in *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, ser. CEAS '11. New York, NY, USA: ACM, 2011, pp. 210–219.
- [16] E. Damiani, S. D. C. di Vimercati, S. Paraboschi, and P. Samarati, "P2p-based collaborative spam detection and filtering," in *Proceedings of the Fourth International Conference on Peer-to-Peer Computing*, ser. P2P '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 176–183.
- [17] F. Li and M. H. Hsieh, "An empirical study of clustering behavior of spammers and group-based anti-spam strategies," in *CEAS 2006 - The Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California, USA*, 2006, pp. 27–28.
- [18] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th Annual Computer Security Applications Conference*, ser. ACSAC '10. New York, NY, USA: ACM, 2010, pp. 1–9.
- [19] D. DeBarr and H. Wechsler, "Using social network analysis for spam detection," in *Proceedings of the Third international conference on Social Computing, Behavioral Modeling, and Prediction*, ser. SBP'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 62–69.
- [20] A. H. Wang, "Don't follow me: Spam detection in twitter," in *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, 2010, pp. 1–10.
- [21] C. X. Jin, X. Lin, J. Luo, and J. Han, "Socialspanguard: A data mining-based spam detection system for social media networks." *PVLDB*, no. 12, pp. 1458–1461, 2011.
- [22] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves, "Detecting spammers and content promoters in online video social networks," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '09. New York, NY, USA: ACM, 2009, pp. 620–627.
- [23] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: web spam detection using the web topology," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '07. New York, NY, USA: ACM, 2007, pp. 423–430.
- [24] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '10. New York, NY, USA: ACM, 2010, pp. 435–442.
- [25] Q. Gan and T. Suel, "Improving web spam classifiers using link structure," in *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, ser. AIRWeb '07. New York, NY, USA: ACM, 2007, pp. 17–20.
- [26] A. Ramachandran, N. Feamster, and S. Vempala, "Filtering spam with behavioral blacklisting," in *Proceedings of the 14th ACM conference on Computer and communications security*, ser. CCS '07. New York, NY, USA: ACM, 2007, pp. 342–351.
- [27] Y. Zhao, Y. Xie, F. Yu, Q. Ke, Y. Yu, Y. Chen, and E. Gillum, "Botgraph: large scale spamming botnet detection," in *Proceedings of the 6th USENIX symposium on Networked systems design and implementation*, ser. NSDI'09. Berkeley, CA, USA: USENIX Association, 2009, pp. 321–334.
- [28] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, ser. IMC '10. New York, NY, USA: ACM, 2010, pp. 35–47.
- [29] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov, "Spamming botnets: signatures and characteristics," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 171–182, Aug. 2008.
- [30] R. Brendel and H. Krawczyk, "Application of social relation graphs for early detection of transient spammers," *WSEAS Trans. Info. Sci. and App.*, vol. 5, no. 3, pp. 267–276, Mar. 2008.
- [31] M. Fire, G. Katz, and Y. Elovici, "Strangers intrusion detection-detecting spammers and fake proles in social networks based on topology anomalies," *Human Journal*, vol. 1, no. 1, pp. 26–39, 2012.
- [32] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. Witten, and L. Trigg, "Weka," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Springer US, 2005, pp. 1305–1314.
- [33] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in Facebook," in *Proc. Workshop on Online Social Networks*, 2009, pp. 37–42.
- [34] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *Proc. European Conf. on Machine Learning*, 2004, pp. 217–226.
- [35] M. Bouguessa, "An unsupervised approach for identifying spammers in social networks," in *Proceedings of the 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, ser. ICTAI '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 832–840.
- [36] A. Lancichinetti and S. Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Physical Review E*, vol. 80, p. 016118, 2009.
- [37] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.
- [38] Y. Freund and L. Mason, "The alternating decision tree learning algorithm," in *Proceedings of the Sixteenth International Conference on Machine Learning*, ser. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 124–133.
- [39] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, Jan. 1991.
- [40] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, ser. UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.