FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 40 (2015) No. 2
ISSN 0867-6356
DE GRUYTER OPEN
DOI: 10.1515/fcds-2015-0006
e-ISSN 2300-3405

CLASSIFIER ENSEMBLES USING STRUCTURAL FEATURES FOR SPAMMER DETECTION IN ONLINE SOCIAL NETWORKS

Muhammad ABULAISH[*] and Sajid Y. BHAT[*]

Abstract. As the online social network technology is gaining all time high popularity and usage, the malicious behavior and attacks of spammers are getting smarter and difficult to track. The newer spamming approaches using the social engineering concepts are making traditional spam and spammer detection techniques obsolete. Especially, content-based filtering of spam messages and spammer profiles in online social networks is becoming difficult. Newer approaches for spammer detection using topological features are gaining attention. Further, the evaluation of ensemble classifiers for detection of spammers over social networking behavior-based features is still in its infancy. In this paper, we present an ensemble learning method for online social network security by evaluating the performance of some basic ensemble classifiers over novel community-based social networking features of legitimate users and spammers in online social networks. The proposed method aims to identify topological and community-based features from users' interaction network and uses popular classifier ensembles – bagging and boosting to identify spammers in online social networks. Experimental evaluation of the proposed method is done over a real-world data set with artificial spammers that follow a behavior as reported in earlier literature. The experimental results reveal that the identified features are highly discriminative to identify spammers in online social networks.

Keywords: Social network security, Spammer detection, Ensemble learning, Classifier ensembles, Feature extraction.

1. Introduction

The Online Social Network (OSN) technology has experienced an exploding popularity and usage in the recent years [20]. This popularity is mostly a result of the numerous features provided by them to their users, which mainly include joining a network (become users), defining their preferences (profile), and publishing any content which they like to share

[*] Department of Computer Science, Jamia Millia Islamia (A Central University), Delhi, India. E-mail: mAbulaish@jmi.ac.in, bhatsajid786@gmail.com

of the interacting nodes in online social networks to identify spammers. Incorporating additional sociological characteristics (like interaction behavior of nodes within and across network community structures) in the classification models can improve their performance for identifying spammers who incorporate sybil and cloning attacks. In [1], we have identified some novel community-based structural features, based on inter- and intra-community users interaction patterns, to learn improved classification models for spammer classification in online social networks. However, the results only spanned over single classifiers and the significance of using ensemble learning methods has not been extensively evaluated yet. In this paper, which is an extended version of our preliminary work on ensemble methods for spammer classification [3], we present details about the topological and community-based features and their significance towards spammer detection in online social networks. We also present an evaluation of individual features to establish their discriminative property for spammer detection using classifier ensembles. Two popular classifier ensembles – *bagging* and *boosting* are used over topological and community-based features extracted from a real-world social network data set with artificially planted spammers. Results are generated for both classifier ensembles using decision tree and naïve Bayes algorithms to evaluate the performance of the proposed spammer detection method.

The rest of the paper is organized as follows. Section 2 presents a brief review of the existing works on spam/spammer detection. Section 3 presents details about the topological and community-based features and their formulations. Section 4 briefly describes two classifier ensembles – bagging and boosting. Section 5 presents the experimental setup and evaluation results. Finally, section 6 concludes the paper with future directions of work.

2. Related work

Spam/Spammer detection methods usually involve two approaches – content-based learning and topology-based learning. In literature, most of the work has been done along the lines of detecting e-mail spam and web spam mainly by exploiting content-based patterns of spam emails and web pages. The main idea behind content-based learning revolves around the observation that spammers use distinguished keywords, URLs, etc. in their interactions and to define their profiles. Such content-based features are used to learn classification models to label messages and profiles as legitimate or spam [31]. However, such approach is often deceived by spammers using copy profiling and content obfuscation.

With the evolution of OSNs the spamming behavior shown and the content disseminated by spammers has changed and got similar to that of legitimate users. This makes detecting spammers in OSNs a highly challenging task. Along these lines topology-based learning methods aim to exploit structural social network features like clustering coefficient, community structures, reciprocity, node degree, etc. to characterize network behavior of legitimate and spammer accounts. Shrivastava *et al.* [30] incorporated features including clustering coefficient and neighborhood independence to deal with random link attacks from spammers. Gan and Suel [16] extracted features like in-links, out-links, cross-links, etc. from a Web graph to classify pages as spam or benign. Other methods include finding physical node clusters based on network-level features from online communication networks [29]. To detect spam clusters, Gao *et al.* [17] used two widely acknowledged distinguishing features of spam campaigns – *distributed coverage* and *bursty nature*. The *distributed* property is quantified using the number of users that send wall posts in the cluster, whereas *bursty* property is based on the intuition that most spam campaigns involve coordinated action by many accounts within short periods of time [33].

In order to, more strictly, characterize the spammers from legitimate users, the methods proposed in [1] and [13] exploit the existence of community structures in social networks [2] to extract novel structural features for the task. Communities resemble groups of nodes that 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. In the context of spammer detection, a node at the boundary of a community which has out links to nodes that belong to other distinct communities may be considered suspicious, as legitimate users tend to show high interactivity within their respective communities [1]. For spammer detection, these methods split the interaction network of OSN users into communities and then extract community-based features of network nodes (users) to classify them as spammer or legitimate.

Besides, there exist ensemble methods that can be used to improve the performance of classifiers by learning multiple models over the same training example set and then using some aggregation method to decide upon a single combined label determined by multiple classifiers. Using ensemble learning methods to improve the performance of spam detection methods have been adopted by many researchers, but their studies have been mainly oriented towards content-based classification of e-mail and web-spam [12]. It is important to note that evaluating ensemble classification methods for spammer detection over topological features of OSN user interaction graphs is still in its infancy. In [18], the authors used an ensemble under-sampling classification strategy incorporating C4.5, bagging, and adaboost. Their results using the ensemble approach showed improvement in Web spam detection performance effectively. Using a text corpus, the authors in [27] aimed to show the significance of classifier ensembles over individual classifiers for spam detection. However, they failed to show any significant improvement in the task. In [22], the authors highlighted the high performance of classifier ensembles involving Adaboost, Stacking, and Ensemble Decision Tree, against the best performance of single classifiers for e-mail spam detection using a content-based approach. In [8], the authors showed that the classifier ensemble proposed by Caruana *et al.* [9] performed better than most individual and classifier ensembles implemented in WEKA for the task of email spam detection. In [12], the authors exploited both content-based and link-based features to compile a minimal feature set that can be computed incrementally in a quick manner to allow intercepting spam. They also showed that for a selected feature set, ensemble classification technique outperforms previously published methods and the Web spam challenge 2008 best results.

3. Topological and community-based features

In this section, we present the formulation of topological and community-based features that are used to learn classifier ensembles for spammer detection in online social networks.

Though topological features for each node (user) in the users' interactions network are defined using graph properties, an overlapping community structure of nodes is identified to define community-based features [4]. The community-based features include the features that 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 a particular node belongs to. Further details about different types of features identified from users' interactions network are given in the following sub-sections.

3.1 Topological features

In this section, various topological features are defined using the basic graph properties. For each feature, a shortened notation is assigned and given in parenthesis for reference in the remaining portion of this paper.

***Total out-degree* (*TOD*):** This is a directed graph-based feature, which captures the interaction behavior of a user with other users in the network. The *total out-degree* of a node (user) *u* represents the total number of distinct users in the social network with whom *u* has direct out-links, i.e., to whom *u* sends messages, etc. It is formally defined using Equation 1, where *V* is the set of nodes and *E* is the set of edges in social networks.

$$TOD(u) = \left|\{v \mid v \in V \wedge (u,v) \in E\}\right| \tag{1}$$

***Total reciprocity* (*TR*):** This is also a directed graph-based feature, which captures the mutual interaction pattern of a user with other users in the network. The *total reciprocity* of a node *u* represents the ratio of the number of mutual interactions of *u* to the total number of nodes with which *u* has out-links. Formally, it can be defined using Equation 2, where L_i^u is the set of links (edges) incident to node *u* and L_o^u is the set of links (edges) originating from *u*.

$$TR(u) = \frac{\left|L_i^u \cap L_o^u\right|}{\left|L_o^u\right|} \tag{2}$$

***Total in/out ratio* (*TIOR*):** This is a more generic feature than the *TR* feature which is defined for a node (user) *u* as the ratio of the number of links (edges) incident to *u* to the number of links (edges) originating from *u*. Formally, it can be defined using Equation 3, where the notations have the same interpretations as given in Equation 2.

$$TIOR(u) = \frac{\left|L_i^u\right|}{\left|L_o^u\right|} \tag{3}$$

nodes with which u has out-links.

$$FIOR(u) = \frac{|F_i^u|}{|F_o^u|} \tag{9}$$

$$F_i^u = |\{v \mid v \in F^u \wedge (v,u) \in E\}| \tag{10}$$

$$F_o^u = |\{v \mid v \in F^u \wedge (u,v) \in E\}| \tag{11}$$

***Foreign reciprocity* (*FR*):** The *foreign reciprocity* of a node *u* represents the ratio of the number of mutual interactions of the node *u* with its foreign nodes to the total number of foreign nodes with whom *u* has out-links. Formally, it is defined using Equation 12.

$$FR(u) = \frac{|F_i^u \cap F_o^u|}{|F_o^u|} \tag{12}$$

***Foreign out-link probability* (*FOLP*):** The foreign out-link probability of a node (user) *u* represents the probability of its out-links to the foreign nodes. It is defined as a ratio of the number of foreign nodes with which *u* has out-links to the total number of nodes with which *u* has out-links, as given in Equation 13.

$$FOLP(u) = \frac{|F_o^u|}{|L_o^u|} \tag{13}$$

***Foreign out-link grouping* (*FOLG*):** The foreign out-link grouping feature of a node *u* represents the probability that the foreign nodes out-linked with *u* have a common community. If $MF_o^u \subseteq F_o^u$ is the maximal set of foreign nodes out-linked with *u* that have a common community, then the *FOLG* feature of *u* is calculated as the ratio of the number of nodes in MF_o^u to the number of nodes in F_o^u, as given in Equation 14.

$$FOLG(u) = \frac{|MF_o^u|}{|F_o^u|} \tag{14}$$

4. Classifier ensembles

Classifier ensembles combine multiple machine learning instances to improve the classification results of a system. It is based on the assumption that combination of multiple

tasks, ranging from data pre-processing and classification rule mining to clustering. Further details about the data set and experimental results are presented in the following sub-sections.

5.1 Data set

As discussed earlier, the approach followed in this paper aims to detect spammers by extracting structural features from the user interaction patterns of OSN users. In this regard the dataset required for analysis is expected to contain a weighted (weights representing the frequency of interactions) network including both legitimate and spammer nodes. However, due to the unavailability of such a dataset and the complexity of extracting the same from a OSN given the access restrictions, we are left with the option of generating an artificial network which would reflect the real world situation to a maximum extent. For the experiments conducted in this paper, we have used a real-world social network data set representing the wall post activities of about 63891 Facebook users [32]. The nodes in this network are considered to be legitimate nodes. We inject additional nodes in the network to simulate 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 network to represent them as legitimate networks. This results in a network with 32693 legitimate nodes. Thereafter, in order to simulate spammers, we generate a set of 1000 isolated nodes for the legitimate network, which creates out-links to randomly selected nodes in the legitimate network. The out-links or the out-degree generated for the spammers are not random, rather it follows the distribution shown by spammers as reported in [19] and used in [6] and [24]. The out-degree distribution of spammer nodes (as reported in [19]) is shown in Table 1. Since the messages of the spammers are expected to be least often reciprocated, the probability of a legitimate node replying to a spammer is set to 0.05.

Table 1. 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

$$TPR = \frac{TP}{TP + FN} \tag{15}$$

$$FPR = \frac{FP}{FP + TN} \tag{16}$$

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$F\text{-}measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{19}$$

Table 3 presents the performance (averaged for both spam and non-spam classes) of the individual classifiers on the data set with planted spammers, wherein it is clear that the decision tree based classifier J48 performs better than the naïve Bayes classifier. Tables 4 and 5 present the performance of the bagging ensemble over the base classifiers – naïve Bayes and J48, respectively, whereas Tables 6 and 7 present the performance of the boosting ensemble over the base classifiers – naïve Bayes and J48, respectively for spammer detection task. It can be observed from these tables that the performance of naïve Bayes and J48 classifiers using either bagging or boosting ensemble learning approach is better than their individual performance. However, in case of naïve Bayes classifier, the ensemble approaches show low performance than the ensemble approaches using J48 classifier.

Table 3. Classification performance (weighted avg. over both classes) of individual classifiers

Classifier	**TPR**	**FPR**	**Precision**	**F-measure**
J48	0.963	0.075	0.963	0.963
Naïve Bayes	0.914	0.175	0.917	0.915

Table 4. Classification performance using bagging with naïve Bayes classifiers

Class	**TPR**	**FPR**	**Precision**	**F-measure**	**ROC area**
Non-spam	0.957	0.175	0.989	0.973	0.959
Spam	0.825	0.043	0.542	0.654	0.959
Weighted avg.	**0.95**	**0.167**	**0.963**	**0.955**	**0.959**

Table 9. Classification performance using bagging with J48 classifiers after excluding one feature at a time

Feature set	Class	TPR	FPR	Precision	F-measure
F - {TIOR}	Non-spam	0.999	0.026	0.998	0.999
	Spam	0.974	0.001	0.979	0.976
	Weighted avg.	0.997	0.025	0.997	0.997
F - {TR}	Non-spam	0.999	0.025	0.999	0.999
	Spam	0.976	0.001	0.98	0.978
	Weighted avg.	0.997	0.023	0.997	0.997
F - {FOLP}	Non-spam	0.999	0.02	0.999	0.999
	Spam	0.98	0.001	0.981	0.981
	Weighted avg.	0.998	0.019	0.998	0.998
F - {FOD}	Non-spam	0.999	0.02	0.999	0.999
	Spam	0.98	0.001	0.981	0.981
	Weighted avg.	0.998	0.019	0.998	0.998
F - {CM}	Non-spam	0.999	0.023	0.999	0.999
	Spam	0.978	0.001	0.981	0.979
	Weighted avg.	0.998	0.021	0.998	0.998
F - {TOD}	Non-spam	0.999	0.026	0.999	0.999
	Spam	0.974	0.001	0.979	0.976
	Weighted avg.	0.997	0.025	0.997	0.997
F - {CN}	Non-spam	0.999	0.022	0.999	0.999
	Spam	0.978	0.001	0.981	0.98
	Weighted avg.	0.998	0.021	0.998	0.998
F - {FR}	Non-spam	0.999	0.022	0.999	0.999
	Spam	0.979	0.001	0.982	0.98
	Weighted avg.	0.998	0.02	0.998	0.998
F - {FIOR}	Non-spam	0.999	0.02	0.999	0.999
	Spam	0.98	0.001	0.981	0.981
	Weighted avg.	0.998	0.019	0.998	0.998
F - {FOLG}	Non-spam	0.999	0.027	0.998	0.999
	Spam	0.974	0.001	0.978	0.976
	Weighted avg.	0.997	0.025	0.997	0.997

Since the performance of ensemble methods using J48 performs better than the ensemble methods using naïve Bayes classifier, and that the performance of bagging with J48 and boosting with J48 is comparable, we have used bagging with J48 classifier for further

References

[1] Bhat S. Y., Abulaish M., Community-based features for identifying spammers in online social networks, in: *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, ACM, 2013, 100-107.

[2] Bhat S. Y., Abulaish M., Analysis and mining of online social networks: emerging trends and challenges, *WIREs: Data Mining and Knowledge Discovery*, 3, 6, 2013, 408-444.

[3] Bhat S. Y., Abulaish M., Mirza A. A., Spammer classification using ensemble methods over structural social network features, in: *Proceedings of the 14th IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Warsaw, Poland, 2014, 454-458.

[4] Bhat S. Y., Abulaish M., HOCTracker: Tracking the evolution of hierarchical and overlapping communities in dynamic social networks, *IEEE Transactions on Knowledge and Data Engineering*, 27, 4, 2014, 1019-1032.

[5] Bilge L., Strufe T., Balzarotti D., Kirda E., All your contacts are belong to us: automated identity theft attacks on social networks, in: *Proceedings of the 18th International Conference on World Wide Web (WWW)*, ACM, NY, USA, 2009, 551-560.

[6] Bouguessa M., An unsupervised approach for identifying spammers in social networks, in: *Proceedings of the IEEE 23rd International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, Washington DC, USA, 2011, 832-840.

[7] Breiman L., Bagging predictors, *Machine Learning*, 24, 2, 1996, 123-140.

[8] Carpinter J. M., *Evaluating Ensemble Classifiers for Spam Filtering*, Honours Thesis, University of Canterbury, 2005.

[9] Caruana R., Niculescu-Mizil A., Crew G., Ksikes A., Ensemble selection from libraries of models, in: *Proceedings of the 21st International Conference on Machine Learning*, 2004, 137–144.

[10] Dietterich T. G., Ensemble methods in machine learning, *Lecture Notes in Computer Science*, 1857, 2000, 1–15.

[11] Douceur J. R., The sybil attack, in: *Revised Papers from the 1st International Workshop on Peer-to-Peer Systems*, Springer-Verlag, London, UK, 2002, 251-260.

[12] Erdélyi M., Garzó A., Benczúr A. A., Web spam classification: a few features worth more, in: *Proceedings of the Joint WICOW/AIRWeb Workshop on Web Quality*, ACM, 2011, 27-34.

[13] Fire M., Katz G., Elovici Y., Strangers intrusion detection-detecting spammers and fake proles in social networks based on topology anomalies, *Human Journal*, 1, 1, 2012, 26–39.

[14] Frank E., Hall M., Holmes G., Kirkby R., Pfahringer B., Witten I., Trigg L. Weka, O. Maimon and L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook*, Springer, 2005, 1305-1314.

[15] Freund Y., Schapire R. E., Experiments with a new boosting algorithm, in *Proceedings of the 13th International Conference on Machine Learning*, 1996, 325-332.

[31] Stringhini G., Kruegel C., Vigna G., Detecting spammers on social networks, in: *Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC)*, ACM, NY, USA, ACM, 2010, 1–9.

[32] Viswanath B., Mislove A., Cha M., Gummadi K. P., On the evolution of user interaction in Facebook, in: *Proceedings of the Workshop on Online Social Networks*, 2009, 37-42.

[33] Xie Y., Yu F., Achan K., Panigrahy R., Hulten G., Osipkov I., Spamming botnets: signatures and characteristics, *SIGCOMM Computing Communication Review*, 38, 4, 2008, 171-182.

Received 7.12.2014, accepted 30.03.2015