A Hybrid Approach for Detecting Automated Spammers in Twitter

Mohd Fazil and Muhammad Abulaish, *Senior Member, IEEE*

***Abstract*— Twitter is one of the most popular microblogging services, which is generally used to share news and updates through short messages restricted to 280 characters. However, its open nature and large user base are frequently exploited by automated spammers, content polluters, and other ill-intended users to commit various cybercrimes, such as cyberbullying, trolling, rumor dissemination, and stalking. Accordingly, a number of approaches have been proposed by researchers to address these problems. However, most of these approaches are based on user characterization and completely disregarding mutual interactions. In this paper, we present a hybrid approach for detecting automated spammers by amalgamating community-based features with other feature categories, namely *metadata-*, *content-*, and *interaction-based* features. The novelty of the proposed approach lies in the characterization of users based on their interactions with their *followers* given that a user can evade features that are related to his/her own activities, but evading those based on the *followers* is difficult. Nineteen different features, including six newly defined features and two redefined features, are identified for learning three classifiers, namely, *random forest*, *decision tree*, and *Bayesian network*, on a real dataset that comprises benign users and spammers. The discrimination power of different feature categories is also analyzed, and interaction- and community-based features are determined to be the most effective for spam detection, whereas metadata-based features are proven to be the least effective.**

***Index Terms*— Social network analysis, spammer detection, spambot detection, social network security.**

I. Introduction

TWITTER, a microblogging service, is considered a popular online social network (OSN) with a large user base and is attracting users from different walks of life and age groups. OSNs enable users to keep in touch with friends, relatives, family members, and people with similar interests, profession, and objectives. In addition, they allow users to interact with one another and form communities. A user can become a member of an OSN by registering and providing details, such as name, birthday, gender, and other contact information. Although a large number of OSNs exist on the web, Facebook and Twitter are among the most popular OSNs and are included in the list of the top 10 websites[1] around the worldwide.

A. OSN and the Social Spam Problem

Twitter, which was founded in 2006, allows its users to post their views, express their thoughts, and share news and other information in the form of tweets that are restricted to 280 characters. Twitter allows the users to follow their favorite politicians, athletes, celebrities, and news channels, and to subscribe to their content without any hindrance. Through *following* activity, a follower can receive status updates of subscribed account. Although Twitter and other OSNs are mainly used for various benign purposes, their open nature, huge user base, and real-time message proliferation have made them lucrative targets for cyber criminals and socialbots. OSNs have been proven to be incubators for a new breed of complex and sophisticated attacks and threats, such as cyberbullying, misinformation diffusion, stalking, identity deception, radicalization, and other illicit activities, in addition to classical cyber attacks, such as spamming, phishing, and drive by download [1], [2]. Over the years, classical attacks have evolved into sophisticated attacks to evade detection mechanisms. A report[2] submitted to the US Securities and Exchange Commission in August 2014 indicates that approximately 14% of Twitter accounts are actually spambots and approximately 9.3% of all tweets are spam. In social networks, spambots are also known as socialbots that mimic human behavior to gain trust in a network and then exploit it for malicious activities [3]. Such reports and findings demonstrate the extent of cyber crimes committed by spambots and how OSNs are proving to be a heaven for these bots. Although spammers are less than benign users, they are capable of affecting network structure and trust for various illicit purposes.

B. Why Connected Users?

Many researchers from academia and industry are working to eliminate the cyber criminals and malicious users to make OSN usage a pleasant and delightful experience. Consequently, a number of spam detection approaches have been proposed. However, as approaches mature and advance,

Manuscript received October 10, 2017; revised January 20, 2018 and March 13, 2018; accepted March 26, 2018. Date of publication April 11, 2018; date of current version May 21, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Anderson Rocha. *(Corresponding author: Muhammad Abulaish.)*

M. Fazil is with the Department of Computer Science, Jamia Millia Islamia, Delhi 110025, India (e-mail: mohdfazil.jmi@gmail.com).

M. Abulaish is with the Department of Computer Science, South Asian University, Delhi 110021, India (e-mail: abulaish@sau.ac.in).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2018.2825958

[1] http://www.alexa.com/topsites

[2] www.techtimes.com/articles/12840/20140812/twitter-acknowledges-14-percent-users-bots-5-percent-spam-bots.html

spammers are using more sophisticated mechanisms to evade detection, thereby resulting in a "cat and mouse game". Fletcher [4] comprehensively analyzed different variants of spammers, starting from conventional spammers to present-day complex spammers, and found that such threats pose dire consequences to different parties associated with the Internet. In addition, Fletcher's paper also discussed the legal challenges related to handling spamming. Boshmaf *et al.* [5] reported that existing spamming and other malicious behavior detection strategies utilize either feature-based or graph partitioning-based strategies. In the first case, users are characterized based on features extracted from their profiles and activities and different classifiers are trained to distinguish between benign users and spammers [6]–[8]. In a feature-based strategy, features, such as number of followers and number of tweets are generally easy to evade, whereas certain complex features are difficult to evade. However, features are generally based on user activities and thus, spammers can regulate their behavior to mimic those of normal users. By contrast, in graph partitioning-based strategies, a user interaction network is partitioned into sub-graphs or communities using graph analysis techniques [9]–[11]. Although, these strategies are formal detection approaches, automated spammers can evade them by creating sufficient attack links (edges) between normal and malicious users. Therefore, we propose an amalgamation of community-based features with other feature categories for detecting automated spammers, wherein communities are identified using graph-partitioning algorithms.

As discussed earlier, most spammer detection approaches are based on the features extracted from user profile and activities in a network. By contrast, spammers advance themselves against these features either by exploiting the loopholes of existing detection techniques or by investing in human or financial resources [12]. Benign users generally follow and respond to requests from known users and avoid connection with and communication from strangers. In this manner, in the network of trust of a user, most users exhibit a certain level of confidence in the identity of others, which leads to the formation of a community-like structure. A benign user may be a member of multiple communities depending on real-world networks and interests. By contrast, spammers generally follow random users, which results in an extremely low reciprocation rate that forms very sparse connections among followers, and adversely affects interaction- and community-based features. To evade features from these categories, spammers may attempt to form a community through mutual following. However, such attempts will be useless because it will not increase their target user base. Consequently, the entire concept of account formation for spamming and maligning is suppressed. Spammers will find bypassing community-based features extremely difficult because the majority of the members of their communities will exhibit spamming behavior which will increase their probability of being exposed.

C. Our Contribution

In this study, we propose a hybrid approach for detecting social spambots in Twitter, which utilizes an amalgamation of *metadata*-, *content*-, *interaction*-, and *community*-based features. In the analysis of characterizing features of existing approaches, most network-based features are not defined using user followers and underlying community structures [6]–[8], [12], [13], thereby disregarding the fact that the reputation of user in a network is inherited from the followers (rather than from the ones user is following) and community members. Therefore, we emphasize the use of followers and community structures to define the network-based features of a user. We classify our set of features into three broad categories, namely, *metadata*, *content*, and *network*, wherein the network category is further classified into interaction- and community-based features. Metadata features are extracted from available additional information regarding the tweets of a user, whereas content-based features aim to observe the message posting behavior of a user and the quality of the text that the user uses in posts. Network-based features are extracted from user interaction network.

The main contributions of this study can be summarized as follows.

- A novel study that uses community-based features with other feature categories, including *metadata*, *content*, and *interaction*, for detecting automated spammers.
- Six new features are introduced and two existing features are redefined to design a feature set with improved discriminative power for segregating benign users and spammers. Among the six new features, one is content-based, three are interaction-based, and the remaining two are community-based. Meanwhile, both redefined features are content-based. When defining interaction-based features, focus should be on the *followers* of a user, rather than on the ones he/she is *followings*.
- A detailed analysis of the working behavior of automated spammers and benign users with respect to newly defined features. In addition, two-tailed Z-test statistical significance analysis is performed to answer the following question: "*is the difference between the working behavior of spammers and benign users in terms of newly defined features a random chance?*"
- A thorough analysis of the discriminating power of each feature category in segregating automated spammers from benign users.

The rest of the paper is organized as follows. Section II presents a brief review of existing state-of-the-art methods for detecting automated spammers. Section III presents our proposed hybrid approach for spammer detection. A formal definition of each identified feature and their extraction process is also provided. Section IV describes the experimental setup and evaluation results. Finally, Section V concludes the study and suggests directions for future work.

II. Related Works

Spams are not new. They have been the source of problems from the very early days of the Internet evolution, during the time of the Advanced Research Project Agency Network (ARPANET) was there and the Internet was still in its infancy state. Spams were reported for the first time in 1978 within

the ARPANET network. During that time, spam was not a serious problem and was not given sufficient attention. Through time, spammers have become sophisticated and have matured, similar to the evolution of email spammers to contemporary socialbots. To deal with this continuously evolving and reconstructing problem, a number of techniques have been proposed and developed by researchers. These techniques target various forms of spammers starting from spam email detection to modern and sophisticated forms of spammers and defaulters, such as socialbots and social spambots. During the early days of spamming, when email systems were the prime victims, Sahami *et al.* [14] proposed textual and non-textual and domain-specific features and learned naive Bayes classifier to segregate spam emails from legitimate ones. Schafer [15] and [16] proposed metadata-based approaches to detect botnets based on compromised email accounts to diffuse mail spams. Spam campaigns on Facebook were analyzed by Gao *et al.* [10] using a similarity graph based on semantic similarity between posts and URLs that point to the same destination. Furthermore, they extracted clusters from a similarity graph, wherein each cluster represents a specific spam campaign. Upon analysis, they determined that most spam sources were hijacked accounts, which exploited the trust of users to redirect legitimate users to phishing sites. In [7] and [8], honey profiles were created and deployed on OSNs to observe the behavior of spammer. Both studies presented different sets of features to discriminate benign users from spammers and evaluated them on different sets of OSNs. Wang [17] used content- and graph-based features to classify malicious and normal profiles on Twitter. In contrast to honey profiles, Wang used Twitter API to crawl the dataset. Yang *et al.* [12], Wang [17], and Ahmed and Abulaish [18], used content- and interaction-based attributes for learning classifiers to segregate spammers from benign users on different OSNs. Yang *et al.* [12] and Ahmed and Abulaish [18] analyzed the contribution of each feature to spammer detection, whereas Yang *et al.* [19] conducted an in-depth empirical analysis of the evasive tactics practised by spammers to bypass detection systems. They also tested the robustness of newly devised features. In [20], Zhu *et al.* used a matrix factorization technique to find the latent features from the sparse activity matrix and adopted social regularization to learn the spam discriminating power of the classifier on the Renren network, one of the most popular OSNs in China. Another spammer detection approach in social media was proposed by Tan *et al.* [21]. This approach emphasizes the original content of genuine users that was hacked by spammers and injected with malicious links to deceive the traditional keyword- and sentence-based spammer detection techniques. The URL is widely exploited by spammers either by injecting it into trending topic tweets or into their own tweets. URLs are generally obfuscated using freely and easily available URL shortening services[3] or Twitter embedded service.[4] URL associated issues were thoroughly observed and analyzed in [13] by proposing a URL-based scheme for detecting spam tweets. The authors analyzed URL redirection chain and extracted a number of features from the chain. In [22], Bhat and Abulaish analyzed the community formation behavior of users and devised community-based features that enlightened the difference between human nature and spammer nature of community formation.

Over time, spammers have evolved to more complex and deceptive variants, such as automated spammers, bots, and political bots, by exploiting various automation techniques. Tools and techniques are being developed everyday and thus, bots can be easily created or hired from third party vendors at extremely low costs. Bots can be used for deceptive, organized, and large scale illicit activities and attacks. On an OSN, bots easily become influential simply by engaging and participating in network activities [23]. A thorough study with a robust and wide range of features, including temporal for analyzing automated spammers, was proposed by Amleshwaram *et al.* [6]. In addition to spambot detection, Amleshwaram *et al.* also track spam campaigns created by spammers. Spammers have changed their tactics and have matured from conventional spamming to spambots to the considerably complex socially engineered bots called socialbots. The experimental proof of the existence of social spambots and the challenges arising due to their presence are discussed in [24].

III. Proposed Approach

From the discussions in the previous section, the features inferred from followers in the interaction feature category and community-based features that are extremely difficult to bypass, have been used in a minimal number of the existing spammer detection methods [19], [22]. Therefore, understanding the theoretical basis of using interaction- and community-based features and describing them in a practical manner is one of the main objectives of the proposed work. Instead of focusing only on individual-centric features, user connections (that form interaction networks) should be analyzed at different levels of granularity for identifying interaction- and community-based features, along the line of the PageRank algorithm [25]. In PageRank, the importance score of a webpage depends on the importance of the incoming webpages, rather than on the outgoing webpages. Thus, referring important webpages by a webpage does not guarantee high importance score for the webpage unless it is not incoming connections by important webpages. A similar approach is applied in our proposed method due to the fact that the stature of user u on a social network is determined based on the user's followers, rather than the following, because the followers of a user cannot be determined by the user.

A. Dataset

For the experimental evaluation of the proposed approach, we use the Twitter dataset provided by [19],[5] which contains 11000 labeled users, including 10000 benign users and 1000 spammers. This dataset also contains the lists of *followers* and *followings* of the labeled users, along with their profile information, such as *username*, *location*, and *userid*.

[3]http://tinyurl.com/,https://bitly.com/

[4]http://t.co/

[5]http://faculty.cse.tamu.edu/guofei/research_release.html

TABLE I
DATASET STATISTICS

Total #users	Labelled users			
	Spammer		Benign	
	#users	#tweets	#users	#tweets
3236467	1000	145096	10000	1209522

TABLE II
SYMBOLS AND THEIR DESCRIPTIONS

Symbol	Description
$\overleftarrow{u}$	Follower set of user u (set of users that follow u)
$\overrightarrow{u}$	Following set of user u (set of users that are followed by u)
N(u)	Total number of tweets tweeted by user u
$u_{\overleftarrow{v}}$	A follower, named v, of user u
$\overrightarrow{u_{\overleftarrow{v}}}$	Following set of the follower v of user u

It also contains tweets and associated details, such as tweet id, tweet time, and favorite count of the labeled users. Table I presents a brief statistics of the dataset provided by [19], where *total #users* includes all the *followers* and *followings* of the labeled benign users and spammers. In this dataset, most of the benign users do not have their list of followers; hence values of their interaction- and community-based features will be zero, which forces classifiers to be biased in spammer detection. Therefore, we consider only instances (128 benign users and 1000 spammers) that have a list of followers, which causes a class imbalance problem. To overcome this problem, we use a state-of-the-art oversampling technique, called the synthetic minority oversampling technique (SMOTE) [26], to generate synthetic samples associated with the minority class of the dataset. For a sample data point in SMOTE, its nearest neighbors are identified and synthetic samples based on the difference between the sample point and its neighbors are generated. A total of 872 instances of the benign class are generated using SMOTE to balance the dataset.

B. Feature Extraction

In the proposed automated spammer detection method, 19 features, including 6 new and 2 redefined features, are identified. The feature set is classified into three broad categories, namely, *metadata-based*, *content-based*, and *network-based* features, based on the types of data used to define a feature. Network-based features are further classified into *interaction-based* and *community-based* features. A brief summary of the features, along with their category and source, is provided in Table III. To the best of our knowledge, features marked as *new* in Table III have not been used in the existing literature for automated spammer detection, whereas features marked as *redefined* redefine existing features and footnotes provide the source of existing features. All 19 features are formally defined in the following subsections. Table II presents the symbols used in feature definitions and their descriptions.

1) Metadata-Based Features: The metadata associated with a file (tweet) represent information components that are used to describe the basic attributes of the file. Metadata can be useful in locating an information source and occasionally proven to be more important than data. In this category, four features are identified and defined in the succeeding paragraphs.

a) Retweet ratio (RR): Automated spammers are not sufficiently intelligent to mimic the tweet-generation behavior of human. To post tweets, bots either retweet the tweets posted by others or generate tweets using probabilistic methods, such as the Markov chain algorithm [28], or tweet from database. Such spamming behavior of spammers can be quantified using RR, which is defined as the ratio of the total number of retweeted tweets to the total number of tweets. Mathematically, it is defined using Equation (1), where *RT(u)* is the number of tweets retweeted by user u. The RR value is expected to be low for benign users and high for spammers.

$$RR(u) = \frac{RT(u)}{N(u)} \tag{1}$$

b) Automated tweet ratio (AR): Manual tweet posting is costly because every account requires a person to operate. Therefore, spamming accounts are programmed using the APIs provided by OSNs. The Twitter API is also public, and it can be easily exploited by spammers to operate multiple accounts for their desired purpose. In the original dataset [19], tweets posted using unregistered third party applications have been considered automated tweets and labeled as *API*. Accordingly, the AR of user u is defined as the ratio of the total number of tweets posted by u using API to the total number of tweets of u. Mathematically, AR is defined using Equation (2), where $A(u)$ is the number of tweets posted by u using API.

$$AR(u) = \frac{A(u)}{N(u)} \tag{2}$$

c) Tweet time standard deviation (TSD): The automation of spammers can be identified through temporal analysis given that they use random number generator algorithms to fix activity time. However, randomization algorithms still follow certain distributions. Bots are programmed to be activated at a specified point in time according to the time activation function. Constraint such as not being active from 11 pm to 2 am, may exist. By contrast, humans are random and unpredictable in their log-in or activity time. Humans are sometimes highly active, whereas they sometimes do not even log-in for days and months. Sometime, they are active during day, and sometimes, at night. The TSD feature is defined to capture the variations in tweet times of a user. Mathematically, it is defined using Equation (3), where t_i is the tweet time of i^{th} tweet, $\bar{t}$ is the mean tweet time, and $N(u)$ is the total number of tweets posted by u. The value of TSD is generally extremely low for automated spammers.

$$TSD(u) = \frac{\sum_{i=1}^{N(u)} (t_i - \bar{t})^2}{N(u)} \tag{3}$$

d) Tweet time interval standard deviation (TISD): Unlike the TSD feature defined earlier, $TISD$ tracks patterns in the time interval of consecutive activities. Bots generally post tweets at regular time intervals using certain random generation algorithms. By contrast, humans exhibit highly

TABLE III
SUMMARY OF IDENTIFIED FEATURES ALONG WITH THEIR CATEGORIES AND SOURCES

Category	Feature	Source
Metadata	Retweet Ratio (*RR*)	[27]
	Automated Tweet Ratio (*AR*)	[19]
	Tweet Time Standard Deviation (*TSD*)	[6]
	Tweet Time Interval Standard Deviation (*TISD*)	[6]
Content	Unique URL Ratio (*UUR*)	[6]*
	Unique Mention Ratio (*UMR*)	[6]*, [7]*
	Content and Hashtag Similarity Ratio (*CHS*)	New
	URL Ratio (*UR*)	[27], [19]
	Mention Ratio (*MR*)	[6], [27], [7]
	Hashtag Ratio (*HTR*)	[27]
	Automated Tweet URL Ratio (*AUR*)	[19]
	Automated Tweet Similarity (*ATS*)	[19]
Interaction	Follower Ratio (*FR*)	New
	Mean Followers Following to Follower Ratio (*MFFFR*)	New
	Follower-based Reputation (*FBR*)	New
	Reputation (*R*)	[6], [8], [19]
	Clustering Coefficient (*CC*)	[19]
Community	Community-based Reputation (*CBR*)	New
	Community-based CC (*CBCC*)	New

* *Already defined in [6], [7], but we have re-defined it.*

irregular behavior. Mathematically, $TISD$ is defined using Equation (4), where $T_1, T_2, \ldots, T_n$ represent the time interval between consecutive tweets and $\bar{T}$ is the mean time interval. The value of $TISD$ for automated spammers is generally extremely low given that they tweet and perform activities at regular intervals.

$$TISD(u) = \frac{\sum_{i=1}^{n}(T_i - \bar{T})^2}{N(u)} \tag{4}$$

2) Content-Based Features: In existing spammer detection methods, content quality has been considered as one of the important indicators of spamming. With time, spammers have evolved by incorporating social engineering and other tactics to evade conventional detection methods that rely on easily evaded characteristics of spamming. During the evolution, tweet quality has been improved. However, when spammers start sending improved quality content, their spamming rate is deprecated, and consequently, their end goal of product or service advertisement is not met. Hence, a trade-off exists between content quality and spamming success rate. However, regardless of all these facts, tweet contents are still used as a helpful parameter to determine the intention of a user. Spammers generally post enticing tweets to deceive users. In the proposed approach, a total number of eight content-based features are identified and defined in the following paragraphs.

a) URL ratio (UR): In Twitter, users generally post their views and thoughts about a topic of interest and share news articles and stories in the form of tweets. These tweets generally include URLs that refer to source pages for detailed information. However, when a user continuously injects URLs into tweets, his/her suspicious intention is reflected. The *UR* of a user u is the ratio of the total number of URLs used in his/her tweets to the total number of tweets posted by u. This feature is highly crucial to spammers because if they do not use URLs in their tweets, then they fail to do what they are supposed to do. The *UR* of user u is mathematically defined using Equation (5), where *U(u)* is the number of URLs used in the tweets of u and *N(u)* is the number of tweets posted by u.

$$UR(u) = \frac{U(u)}{N(u)} \tag{5}$$

Spammers use URLs in most of their tweets; hence their *UR* value may approach to 1, and occasionally, even greater than 1. For benign users, the *UR* value is very small (near 0), thereby verifying the fact that the tweets of genuine users are either quotes, thoughts, or views about a topic of interest.

b) Unique URL ratio (UUR): The excessive embedding of URLs in tweets is generally suspicious, but if same URL is used repetitively in tweets, then the user posting the tweets is placed in a highly suspicious category. Spammers generally use the same URL repeatedly in their tweets with the intention that users will be trapped and click on the URL that will redirect them to a malicious site and become a victim of malware attack. Such spamming behavior is observed using a unique URL ratio that captures the uniqueness among the URLs used in the tweets by the user. The *UUR* of user u is calculated using Equation (6), where *UU(u)* is the number of unique URLs and *U(u)* is the number of URLs used in the tweets of u.

$$UUR(u) = \frac{UU(u)}{U(u)} \tag{6}$$

On the basis of the preceding discussions, the *UUR* value will be generally low for spammers and high for benign users.

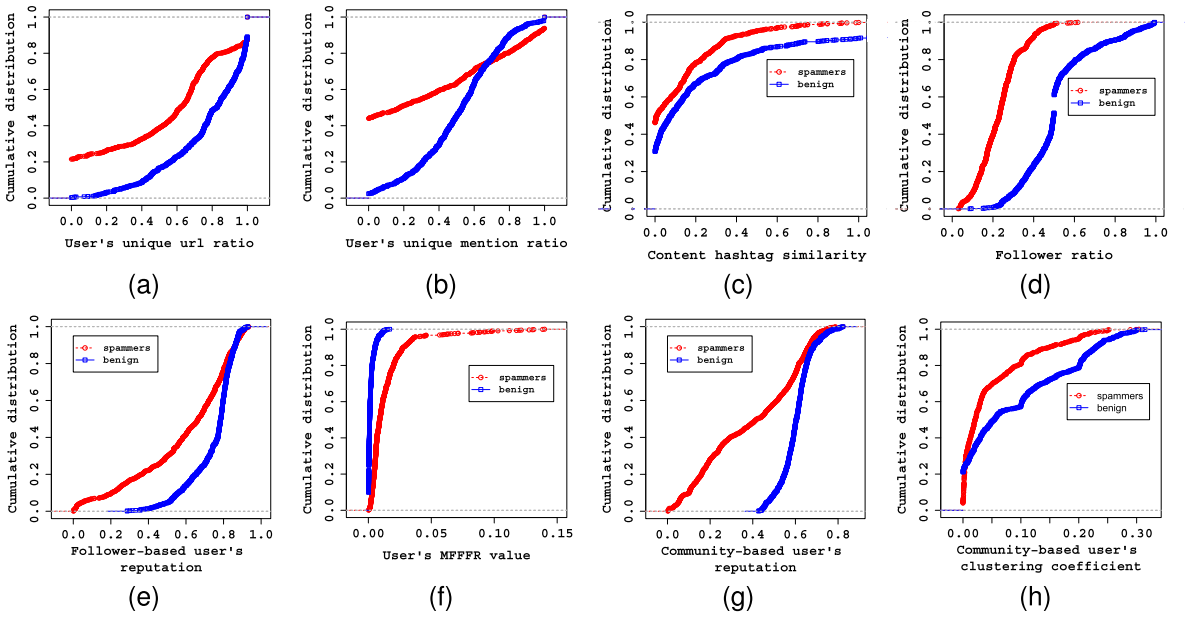

Fig. 1. Cumulative distribution of newly defined features with respect to the dataset described in Section III-A.

However, most spammers are advertisers and use shortened URLs in tweets. To bypass this feature, spammers have to generate different shortened URLs using various URL shortening services, which is infeasible because for an input URL, URL shortening services create the same shortened URL even if it is submitted multiple times without any change. To observe *UUR* among the users, its cumulative distribution plot is shown in Figure 1(a). As indicated in the figure, spammers can generate different shortened URLs for the same URL until a certain extent by using either different shortening services or other techniques, but *UUR(u)* can still be used as a vital feature to segregate spammers from benign users.

c) Mention ratio (MR): Users can be tagged in a tweet using the "@" symbol followed by their Twitter handles. This feature is also abused by spammers who mention users in tweets, thereby provoking and enticing them to know about the sender of the message. This self-indulgent nature makes benign users susceptible victims. For example, a user has to mention another user John who is unknown to the user. The user can inscribe John in a tweet as "@" followed by John's Twitter handle, thereby sending a notification to John informing him about the user's tweet. The *MR* for user u is calculated using Equation (7), where *M(u)* is the number of mentions in the tweets and *N(u)* is the number of tweets posted by u.

$$MR(u) = \frac{M(u)}{N(u)} \tag{7}$$

In general, the value of the MR feature is low for benign users and high for spammers.

d) Unique mention ratio (UMR): Genuine users generally have connections with a number of people, including friends, family members, and colleagues, but their interactions happen with only a small subset of persons. By contrast, spammers tag users randomly either from a set of connected users or users from outside of their network of trust. This spamming behavior can be used to segregate spammers from benign users. Mathematically, *UMR* for user u is defined using Equation (8), where *UM(u)* is the number of unique mentions used by u.

$$UMR(u) = \frac{UM(u)}{M(u)} \tag{8}$$

Evidently, the value of *UMR* is low for genuine users given that they interact with a selected set of people, whereas it is high for spammers. If spammer targets a specific set of users by continuously mentioning them in tweets, then the spammer will be noticed and reported by the user, and the malicious intention of the spammer will be exposed. To observe the behavior of the two classes of users, the cumulative distribution of *UMR* is plotted as shown in Figure 1(b). As indicated in the figure, spammers significantly differ from benign users in terms of *UMR*.

e) Content and hashtag similarity (CHS): Twitter lists the top 10 most frequently used hashtags, aka trending topics, on a section of user's wall. Spammers exploit the trending topics listed by Twitter by injecting them into their malicious tweets. Although spammers inject trending hashtags into their tweets, these hashtags and tweets content have no semantic relation. As a result of hashtag injection, whenever benign user search tweets that correspond to a trending hashtag, malicious tweets injected with that hashtag will also be displayed in the searched result, thereby raising the probability of the user becoming a victim of spamming. The *CHS* feature is defined to capture such typical social engineering tactics used by spammers. Mathematically, it is defined using Equation (9), where $MH^i(u)$ is the number of words that match with the hashtags used in the i^{th} tweet of user u, $HT^i(u)$ is the number of hashtags used in the i^{th} tweet, and $N(u)$ is the total number

of tweets by u.

$$CHS(u) = \frac{\sum_{i=1}^{N(u)} \frac{MH^i(u)}{HT^i(u)}}{N(u)} \tag{9}$$

The value of *CHS* is generally high for benign users given that their hashtags and tweet topics are generally the same, and close to 0 for spammers. The cumulative distribution plot of *CHS* values is shown in Figure 1(c), wherein spammers and benign users do not considerably differ in terms of *CHS* and 80% of spammers have a *CHS* value less than 0.2.

f) Hashtag ratio (HTR): Hashtags in Twitter are used to group tweets related to any topic of discussion. Unlike other OSNs, such as Facebook, where groups are created with the name representing the topic of interest. A group for discussing a topic of interest is created in Twitter through hashtags. Twitter displays a list of the top 10 trending hashtags at any moment. This feature is exploited by spammers by hijacking these trending topics. Spammers inject popular hashtags into their tweets, such that whenever these hashtags are searched, tweets by the spammers that contain the searched hashtags are also shown in the search result. The *HTR* for user u is defined using Equation (10), where *HT(u)* is the number of hashtags used in the tweets and *N(u)* is the total number of tweets posted by u. In general, the value of *HTR* for spammers is high, whereas it is low for benign users.

$$HTR(u) = \frac{HT(u)}{N(u)} \tag{10}$$

g) Automated tweet URL ratio (AUR): To capture content quality in the automated tweets of users, this feature is highly important because it analyzes the use of URLs in automated tweets. The *AUR* of user u is defined as the ratio of the number of automated tweets with URLs to the total number of automated tweets by u as defined using Equation (11), where *AU(u)* is the total number of automated tweets with URLs posted by u and *A(u)* is the number of automated tweets by u.

$$AUR(u) = \frac{AU(u)}{A(u)} \tag{11}$$

h) Automated tweet similarity (ATS): This feature aims to find similarity among the automated tweets posted by users given that spammers are generally inclined toward a topic controlled by polluters and fraudsters. Thus, they keep on posting similar contents repeatedly. This feature for user u analyzes content behavior among the automated tweets of u and is calculated using cosine similarity as defined using Equation (12), where $A(u)$ is the number of automated tweets by u and $AU^i(u)$ is the word vector that corresponds to the i^{th} automated tweet. Cosine similarity is calculated between every pair of automated tweets using their word vector. For example, a pair of tweets $A^i(u)$, $A^{i+1}(u)$, is assumed to find cosine similarity. First, both tweets are split using blank space (" ") as delimiter and stop words are filtered. Thereafter, a vector that includes all the words from both tweets is created. Subsequently, a separate vector that corresponds to each tweet of the pair is created, with a value of 1 if the word is present in the tweet and 0 otherwise. Finally, the cosine similarity for the tweet pair is calculated using Equation (12).

$$ATS(u) = \frac{2 \times \sum_{i=1}^{A(u)} \sum_{j=i+1}^{A(u)} \frac{AU^i(u).AU^j(u)}{||AU^i(u)||.||AU^j(u)||}}{A(u)(A(u)-1)} \tag{12}$$

3) Interaction-Based Features: The interaction data that are available from virtual environment through OSNs are rich knowledge source that can be used in intelligent decision-making, such as fraud detection, customer behavior analytics, the real-world identity and behavior prediction of a user. Twitter, an open nature OSN, permits a user to follow other users to subscribe to their tweets and activities, but a user cannot force others to follow him/her back. This nature that allows a user to connect with other users creates a network of trust among the users. Five interaction-based features are identified and discussed in the following sections.

a) Follower ratio (FR): In Twitter, the number of followers of a user generally indicates the trust level of the user among the users of the network. In case of genuine users, the users in the network of trust generally know each other in the real world, except for celebrities and popular users. Therefore, genuine users generally have a high follow-back rate, which can be used to label a user as either spammer or benign. The *FR* of user u represents the follower fraction in the network of trust. Mathematically, it is defined as the ratio of the number of followers gained by user u to the total number of users connected to u as represented using Equation (13).

$$FR(u) = \frac{|\overleftarrow{u}|}{|\overrightarrow{u} \bigcup \overleftarrow{u}|} \tag{13}$$

The value of FR is generally high for genuine users and low for spammers. To observe the difference in the connection-forming behavior between benign users and spammers in terms of *FR*, its cumulative distribution for the two classes is plotted as shown using Figure 1(d). As indicated in the figure, approximately 80% of benign users have a *FR* value higher than 0.4, which is approximately 10% in the case of spammers.

b) Reputation (R): In the real world, the reputation of a user within a society or organization reflects the views and trusts of community users regarding the user. This assumption also holds true in the virtual world. In the context of Twitter, this assumption implies that if user u follows another user v, then the probability that v will follow back u is high, which increases the reciprocity rate of u. The reciprocity rate for user u is the fraction of the users in the network of trust who follow back the user in response to his/her followings. That is, it is the response rate of the connection request sent by a user in the network of trust. Sophisticated spammers bypass this feature, either by mutually following each other or obtaining followers from follower-selling vendors. The R of user u is defined using Equation (14), where $\overrightarrow{u}$ is the set of *followings* and $\overleftarrow{u}$ is the set of followers of u.

$$R(u) = \frac{|\overleftarrow{u} \bigcap \overrightarrow{u}|}{|\overrightarrow{u}|} \tag{14}$$

The R value of spammers is generally low due to the low response from *followings*, whereas that for benign users is high because they generally follow known users except for celebrities.

c) Follower-based Reputation (FBR): The reputation of users is generally not their own but inherited from connected users. In a network, society, or organization, the reputation of a user depends on the users who are within the proximity, and followers are the most significant among them. In general, users have no control over their followers [29] and features based on followers are difficult to evade and tamper. The reputation of a user is directly proportional to the reputation of his/her followers. This feature captures the reflection of the reputation of the followers of a user. The *FBR* of user u is the average of the reputation of the followers of u. Mathematically, it is defined using Equation (15), where $\overleftarrow{u}$ is the follower set of u and $R(u_{\overleftarrow{v}})$ is the reputation of a follower $u_{\overleftarrow{v}}$, which is calculated using Equation (14).

$$FBR(u) = \frac{\sum_{u_{\overleftarrow{v}} \epsilon \overleftarrow{u}} R(u_{\overleftarrow{v}})}{|\overleftarrow{u}|} \tag{15}$$

The value of *FBR* for a benign user should be high given that most of his/her followers will also be benign, whereas its value for spammers is expected to be low. The cumulative distribution based on *FBR* for benign users and spammers is plotted in the Figure 1(e). As shown in the figure, approximately 40% of spammers have *FBR* values less than 0.6, whereas the value for benign users is approximately 10%.

d) Mean follower's followings to followers ratio (MFFFR): To inspect a user, his/her connecting or interacting persons should be examined. Unlike spammers who are very responsive to every request regardless of the sender's identity simply to increase their list of followers, benign users are conscious when responding to request from unknown users. Therefore, to examine the connecting behavior of the followers of a user, we analyze the *following* patterns of the followers with the follower patterns of the user. The *MFFFR* of user u is defined as the ratio of the mean of the follower's *following* to the total number of followers of the user as represented using Equation (16), where $\overleftarrow{u}$ is the follower set of u, $u_{\overleftarrow{v}}$ is one of the followers of u, and $\overrightarrow{u_{\overleftarrow{v}}}$ is the *following* set of the follower $u_{\overleftarrow{v}}$.

$$MFFFR(u) = \frac{\left(\sum_{u_{\overleftarrow{v}} \epsilon \overleftarrow{u}} |\overrightarrow{u_{\overleftarrow{v}}}|\right) / |\overleftarrow{u}|}{|\overleftarrow{u}|} \tag{16}$$

Most of the followers of spammer are also possible spammers or content polluters, and thus, they blindly follow users, thereby leading to the large value of followers' average followings, but their own follower count is less, i.e., the *MFFFR* value for spammers is very high, whereas it is moderate for benign users because they have a moderate number of followers, and the followers themselves have a moderate number of *followings*. These facts can be verified from the cumulative distributions shown in Figure 1(f). On average, spammers have a higher value of *MFFFR* than benign users.

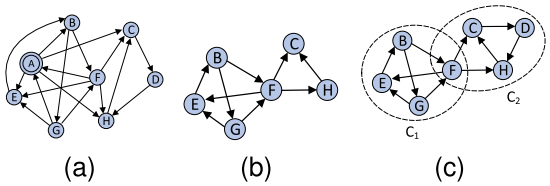

Fig. 2. (a) Followings/followers network, (b) inter-connection network of the neighbors of user "A", and (c) sample communities extracted from the interaction network of the neighbors of user "A".

e) Clustering coefficient (CC): The *CC* of a node in a network represents how densely its adjacent nodes are connected among themselves, excluding the connections that originate from and that terminate at the node. It represents the trust level among the connecting nodes of a user. The closer the value of *CC* of a user to 1, the higher the trust level among the connecting nodes of the user. Considering a small network of followings and followers of a user "A" as shown in Figure 2(a), neighbors of "A" and their inter-connections (nodes that are directly connected with "A") in Figure 2(b), the *CC* of "A" can be calculated as the ratio of number of existing edges in Figure 2(b) to the total number of possible edges among the nodes in Figure 2(b). Formally, for a user u, if K_u and E_u represent the number of connected users (neighbors) and the number of existing edges among the connected users, respectively, then the *CC* of u, $CC(u)$, is calculated using Equation (17).

$$CC(u) = \frac{E_u}{K_u \times (K_u - 1)} \tag{17}$$

For a legitimate user, *CC* is generally high given that the connected nodes of the legitimate user either know one another or have certain level of confidence in the identity of others through trust propagation, thereby leading to the formation of a moderately dense network compared with a spammer network. By contrast, bots or spammers have no real-world identity, which raises suspicions among the followed users.

4) Community-Based Features: From ancient time human beings used to live in communities and societies. This very nature of human is also reflected in virtual world, especially in OSNs. In communities, either user know one another or exhibit a certain level of trust among themselves and have higher connection density among themselves, in comparison to users from outside the communities [30]. In social network, it is generally formed on the basis of similar interest, location, profession, or some other cohesiveness. Formally, a community structure can be defined as a tightly-knit group of vertices with high intra-group edge density and low inter-group edge density [31]. Finding such groups has lot of real-life applications, such as identifying groups of similar entities in social networks, biological networks, citation networks, transportation networks, and e-mail networks. In this paper, we consider communities based on the connections of users arising because of their *followers* and *followings*. We have used CONCLUDE [32], a fast community detection algorithm from large networks, to identify communities in users interaction

network. Figure 2 presents a exemplary users interaction network. Figure 2(c) presents the neighbors of user "A" and their connections with other users, in which C_1 and C_2 represent possible community structures. For spammers, users in the network of trust hardly know one another because of random *followings* and *followers*, thereby leading to the formation of very sparse or no community. Thus, behavior of user can also be determined by analyzing the members of their communities. Therefore, in this paper, we have considered two community-based features that are defined in the following sections.

a) Community-based reputation (CBR): The reputation of a user depends on the reputation of communities and its member, to which user is associated. For a genuine user, member of his/her associated communities exhibit good reputation that eventually increases the reputation of the user. This phenomenon is unlikely to be true in the case of spammers. To calculate the value of *CBR* for user u, first we find communities from the neighbors network of u. Thereafter, calculate the reputation of every user of each community, and eventually, reputation of each community based on the reputation of its member. Finally, reputation of u is averaged over all his/her associated communities. For example, if user u is the member of k communities, say $C_1, C_2, \ldots, C_k$, then the *CBR* value of u is calculated using Equation (18), where C_i represents the i^{th} community of u and $R(C_i(j))$ represents the reputation of the j^{th} user of the i^{th} community of u.

$$CBR(u) = \frac{\sum_{i=1}^{k}\left(\left(\sum_{j=1}^{|C_i|} R(C_i(j))\right) \big/ |C_i|\right)}{k} \tag{18}$$

The *CBR* value for spammers is generally low, and it is high for benign users. To observe this effect, the cumulative distribution of *CBR* values of spammers and benign users is plotted in Figure 1(g). As indicated in the figure, approximately 90% of the benign users have a *CBR* value greater than 0.5, which is only 30%, in the case of spammers.

b) Community-based clustering coefficient (CBCC): This feature studies the connection density among the community members of a user. Suppose, user u is the member of k communities, say $C_1, C_2, \ldots, C_k$ and CC_i is the clustering coefficient of the i^{th} community, which is calculated using Equation (17), then *CBCC* of u is mathematically defined using Equation (19).

$$CBCC(u) = \frac{\sum_{i=1}^{k} CC_i}{k} \tag{19}$$

The value of *CBCC* is generally high for benign users and low for spammers, given that the user connected to spammers have very sparse connections among themselves, and thus, hardly create communities, thereby leading to low value of $CBCC$. As indicated in Figure 1(h), approximately 60% of the benign users have *CBCC* values greater than 0.05, which is approximately 30%, in the case of spammers, thereby establishes the fact that benign users are the member of dense communities and spammers are the members of sparse communities.

IV. Experimental Setup and Results

Following a detailed description of the features and feature extraction process in preceding sections, this section presents the experimental details and evaluation results of the proposed approach for detecting automated spammers in Twitter. The performance of the proposed study is analyzed using three machine learning techniques, namely *random forest*, *decision tree*, and *Bayesian network* on the dataset described in Section III-A. The source code of the proposed method is uploaded at Git-Hub, which can be accessed using the URL: https://goo.gl/cTqiUp.

A. Evaluation Metrics

The proposed approach is evaluated using three standard metrics, namely, *detection rate* (*DR*), *false positive rate* (*FPR*), and *F-Score*. *DR* (aka *recall*) represents the fraction of spammers detected from the set of all spammers, and it is defined using Equation (20), where *TP* stands for true positives and represents the number of actual spammers classified as spammers, and *FN* stands for false negatives and represents the number of actual spammers misclassified as benign users. *FPR* is false positive rate and represents the fraction of benign users, misclassified as spammers, and it is defined using Equation (21), where *FP* stands for false positives and represents the number of benign users misclassified as spammers and *TN* stands for true negatives and represents the number of benign users classified as benign. *FPR* is crucial parameter for evaluation of classifiers, and its low value is desirable for good classifier. Finally, *F-Score* is defined as the harmonic mean of *precision* and *recall* as given in Equation (22), where *precision* is defined as the ratio of the correctly identified spammers to the total number of users identified as spammers, and *recall* is same as the *DR*. The *F-Score* represents discriminative power of classifier. A classifier with a high value of *F-Score* is desirable to precisely segregate the spammers and benign users.

$$DR = \frac{TP}{TP + FN} \tag{20}$$

$$FPR = \frac{FP}{FP + TN} \tag{21}$$

$$F\text{-}Score = \frac{2 \times precision \times recall}{precision + recall} \tag{22}$$

B. Evaluation Results

As stated in the beginning of this section, the performance of the proposed approach is evaluated using three classifiers, namely, *random forest*, *decision tree*, and *Bayesian network*, which are implemented in Weka.[6] We have used ten-fold cross validation to ensure the participation of each instance in both training as well as testing procedure. The performance of the classifiers is evaluated using standard evaluation metrics, namely, *DR*, *FPR*, and *F-Score* that are defined in Section IV-A. The experimental results on various feature categories using the dataset presented in Section III-A is given

[6]http://www.cs.waikato.ac.nz/ml/weka/

TABLE IV
PERFORMANCE EVALUATION OF CLASSIFIERS OVER THE DATASET PRESENTED IN SECTION 3.1

Feature Set	Random Forest			Decision Tree			Bayesian Network		
	DR	FPR	F-score	DR	FPR	F-score	DR	FPR	F-score
F	0.976	0.017	0.979	0.949	0.047	0.943	0.908	0.019	0.942
F\Metadata Feature Set	0.965	0.031	0.972	0.933	0.056	0.949	0.924	0.026	0.948
F\Content Feature Set	0.956	0.028	0.964	0.924	0.057	0.933	0.906	0.041	0.947
F\Interaction Feature Set	0.930	0.027	0.950	0.936	0.053	0.938	0.850	0.046	0.897
F\Community Feature Set	0.949	0.023	0.956	0.931	0.056	0.938	0.843	0.022	0.904

in Table IV. The first row of this table presents the evaluation results of the classifiers considering all 19 features (F). It can be observed from the first row of this table that *random forest* performs best in terms of all three metrics *DR*, *FPR*, and *F-Score*. However *decision tree* is also good in terms of *DR* and *F-Score* with the values of 94.9% and 94.3% respectively. *Bayesian network* performs significantly good in terms of *FPR* and *F-Score*, but not as good in terms of *DR*.

To evaluate the discriminative power of features categories, we perform *feature ablation test* in which some features are removed from the feature set to observe their impact on classifiers' performance [33]. Accordingly, the experiment mentioned above is repeated four times, excluding the features of a particular category in each repetition, using the set-difference function given in Equation (23), where F is the set of all 19 features and F_1 is the set of features of a particular category. The second, third, fourth, and fifth rows of Table IV present the evaluation results corresponding to the exclusion of feature categories. As presented in the table, overall, interaction-based features are efficient in terms of *DR* and *F-Score*. This feature category includes three new features and all the three are based on followers of user, which is one of the novelties of the proposed approach, which reflects the importance of followers for detecting spammers. Content-based features also show moderate discriminating power for *decision tree*, although not good for the other two classifiers, that endorse the fact that bots still use content to trap users by using enticing contents in their posts and it does not depend on their sophistication level. As observed from the table that community-based features also show good discriminating power, and affect the classifiers efficiency. In addition, community-based features are the most discriminating features in terms of *DR* for *Bayesian network*. Metadata features show least impact on performance of the classifiers, which highlights the efficacy of random number generator algorithms, used by bots to achieve randomness in their behavior similar to those of human-beings.

$$F \setminus F_1 = \{x | x \in F \wedge x \notin F_1\} \tag{23}$$

C. Discussion

In cyberspace, there are vendors that lets you gain followers at very cheap cost,[7] such as 18000 followers for $15 only. Spammers generally manage followers through mutual consent by following each other to evade network related features. In the analysis of user profiles, spammers are found to be successful in gaining followers and have approximately same number of followers as of benign users. This finding indicates that conventional features, such as number of followers, followers to followings ratio, retweet ratio based on user data are insufficient for development of effective spammer detection systems. By contrast, community-, follower-, and content-based features, defined using user and his/her followers data enable the development of effective spammer detection systems. Although metadata-based features are generally least contributing in spammer detection due to the fact that spammers exploit randomization algorithms to imitate randomness of human-beings. The *retweet ratio* is found highly variable for benign users given that they generally retweet the tweets of others. The experimental results shown in the preceding sections reflect that newly defined interaction-based features based on followers are highly effective in detecting social spambot. Interaction-based features are also balanced in terms of both *FPR* and *F-Score* values. In addition, community-based features also show significant performance and aid in the improvement of the classifiers accuracy. In summary, it is concluded that followers and communities from the interaction network of user are strong indicators of the reputation of users. Features based on these categories can be used for efficient segregation of social spambots and benign users, in contrast to conventional user-centric features.

1) Results on the Datasets With Varying Spammers and Benign Users Ratio: This section analyzes the effect of different ration of spammers and benign users on the evaluation results. To this end, we repeated the experiment for different ration of spammers and benign users viz 1:1, 1:2, 1:5, and 1:10 and presented the evaluation results in terms of the three evaluation metrics, namely, *DR*, *FPR*, and *F-Score* for the classifiers – *random forest*, *decision tree*, and *Bayesian network* in Table V. The dataset with spammers and benign users ration as 1:1 has 1000 spammers and 1000 benign users, whereas dataset with 1:2 ration has 500 spammers and 1000 benign users. In the analysis, dataset ratio shows correlation with the efficiency metrics except the *Bayesian network* where the correlation is not significant and ratio of spammers and benign users in the dataset does not show significant effect on the result. In the case of *random forest* and *decision tree*, as we decrease the ratio of spammers in the dataset, accuracy of the classifiers also decreases, which is highly significant in case of *random forest*. In this case, the value of *DR* significantly decreases on decreasing the

[7] www.buycheapfollowersfast.com/twitter/

TABLE V
PERFORMANCE EVALUATION OF THE PROPOSED METHOD ON DATASETS WITH DIFFERENT RATIO OF SPAMMERS AND BENIGN USERS

Spammers and	**random forest**			**decision tree**			**Bayesian network**		
Benign Users Ratio	DR	FPR	F-score	DR	FPR	F-score	DR	FPR	F-score
1:1	0.976	0.017	0.979	0.949	0.047	0.943	0.908	0.019	0.942
1:2	0.960	0.009	0.965	0.926	0.027	0.935	0.888	0.016	0.929
1:5	0.921	0.007	0.927	0.895	0.023	0.882	0.885	0.016	0.901
1:10	0.870	0.002	0.919	0.860	0.013	0.864	0.901	0.005	0.935

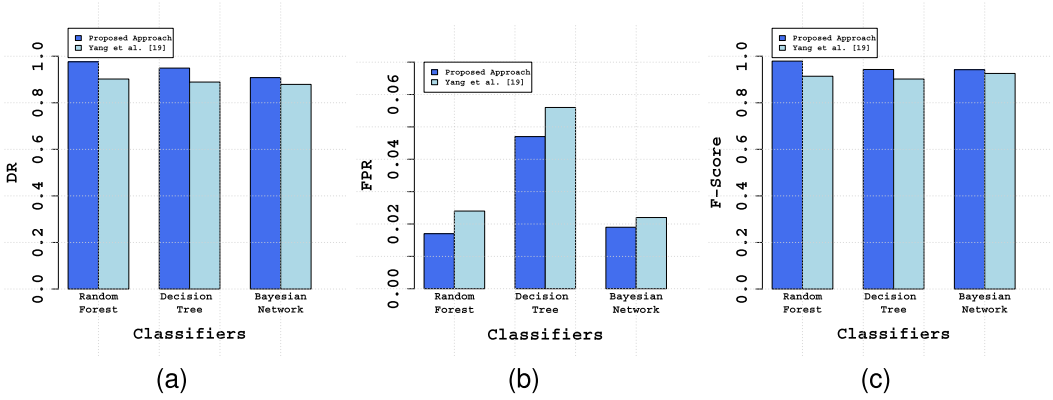

Fig. 3. Performance comparison results over the dataset presented in Section III-A.

spammers in the dataset as shown in the Table V. This decrease in the performance of the classifiers may be due to class imbalance problem.

2) Statistical Significance Analysis of the Behavioral Difference of Spammers and Benign Users: This section presents a statistical significance analysis to answer the question: "*is the difference between the working behavior of spammers and benign users in terms of newly defined features a random chance?*" To test this hypothesis, we perform two-tailed Z-test using the dataset discussed in Section III-A. In Z-test, under null hypothesis, test statistics is supposed to follow the normal distribution, and is evaluated using two hypotheses – *null hypothesis* ($H_0 : \mu = \mu_0$) and *alternative hypothesis* ($H_1 : \mu \neq \mu_0$) [34].

In null hypothesis, assumption is that there is no significant difference in the population means of newly defined feature values between spammers and benign users, whereas in alternative hypothesis, we assume that means of the feature values of the two classes differ significantly. Thereafter, test statistics for each of the 6 newly defined and 2 redefined features is calculated and compared with the tabulated critical values of two-tailed Z-statistics at 5% significance level, which is ± 1.96. In the analysis, null hypothesis for all the eight features is rejected as shown in Table VI. Therefore, it is concluded that mean values of all the eight features for spammers and benign users differ significantly. As evident from the table, mean values of *FR* and *CBCC* are the most significantly different between spammers and benign users. Therefore, based on the cumulative distributions plotted in Figure 1 and significance analysis performed in this subsection, it is inferred that newly defined features are significant to segregate spammers and benign users.

TABLE VI
Z-TEST STATISTICS FOR 6 NEWLY DEFINED AND 2 REDEFINED FEATURES

Feature	Z-test value	$H_0 : \mu = \mu_0$
UUR	2.650	Rejected
UMR	3.127	Rejected
CHS	3.284	Rejected
FR	8.574	Rejected
FBR	6.856	Rejected
MFFFR	4.265	Rejected
CBR	6.213	Rejected
CBCC	7.024	Rejected

D. Comparative Analysis

This section presents a comparative analysis of the proposed method with one of the state-of-art methods for detecting automated spammer proposed by Yang *et al.* [19]. The approach presented in [19] is implemented and evaluated on the dataset discussed in Section III-A and compared with our proposed method using the same set of classifiers. Figure 3 presents the performance comparison of the proposed method with Yang et al. method in terms of *DR*, *FPR*, and *F-Score* respectively. It can be observed from this figure that the proposed approach outperforms Yang et al. method in terms of *DR* for all three classifiers, and *random forest* performs significantly better than the other two classifiers. Similarly, in terms of *FPR* too, the proposed approach outperforms [19] for all the three classifiers and difference is significant for *decision tree*, as shown in the Figure 3(b). Finally, in terms of *F-Score* too the proposed approach outperforms [19] for all three classifiers. Since the spammers detection method reported in [19] has already been

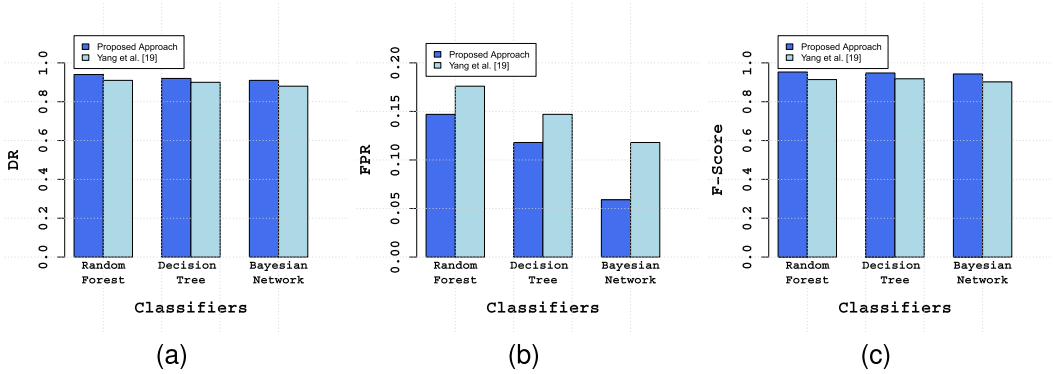

Fig. 4. Performance comparison results over the dataset presented in Section IV-D1.

compared with other state-of-the-art methods reported in [7], [8], [27], and [35], and it has shown comparatively better performance, we have not compared our proposed method with these methods.

1) Comparision on a Relatively Balanced Dataset: In this section, we have performed another experiment over a relatively balanced dataset to further strengthen the efficacy of the proposed method. As discussed in Section III-A, the original dataset provided by Yang *et al.* [19] contains complete information about 1000 spammers and only 128 benign users, which is largely unbalanced. Therefore, we have crawled the profiles of 196 benign users[8] from Twitter, including their personal information, followers, followings, and tweets. Users are labelled benign because either one of the authors knows them in real-world or interacted with them in the past. As a result, the dataset contains total 1000 spammers and 324 benign users. We selected 10% users (100 spammers and 34 benign users) from both categories as test set, and remaining dataset is balanced using SMOTE for training purpose. Thereafter, all three classifiers are trained using our proposed method and Yang *et al.* [19] method, and validated over the test set. Figure 4 presents the performance evaluation results of both methods on this relatively balanced dataset.

V. Conclusion and Future Work

In this paper, we have proposed a hybrid approach exploiting community-based features with *metadata-*, *content-*, and *interaction-based* features for detecting automated spammers in Twitter. Spammers are generally planted in OSNs for varied purposes, but absence of real-life identity hinders them to join the trust network of benign users. Therefore, spammers randomly follow a number of users, but rarely followed back by them, which results in low edge density among their *followers* and *followings*. This type of spammers interaction pattern can be exploited for the development of effective spammers detection systems. Unlike existing approaches of characterizing spammers based on their own profiles, the novelty of the proposed approach lies in the characterization of a spammer based on its neighboring nodes (especially, the followers) and their interaction network. This is mainly due to the fact that users can evade features that are related to their own activities, but it is difficult to evade those that are based on their followers. On analysis, metadata-based features are found to be least effective as they can be easily evaded by the sophisticated spammers by using random number generator algorithms. On the other hand, both interaction- and community-based features are found to be the most discriminative for spammers detection.

Attaining perfect accuracy in spammers detection is extremely difficult, and accordingly any feature set can never be considered as complete and sound, as spammers keep on changing their operating behavior to evade detection mechanism. Therefore, in addition to profile-based characterization, complete logs of spammers starting from their entry in the network to their detection, need to be analyzed to model the evolutionary behavior and phases of the life-cycles of spammers. But, generally spammers are detected when they are at very advanced stage, and it is difficult to get their past logs data. Moreover, it may happen that a user is operative in the network as a benign user, and later on, it starts illicit activities due to whatsoever reasons, and considered as spammer. In this circumstance, even analyzing log data may lead to wrong characterization.

Analysis of spammers network to unearth different types of coordinated spam campaigns run by the spambots seems one of the promising future directions of research. Moreover, analyzing the temporal evolution of spammers' followers may reveal some interesting patterns that can be utilized for spammers characterization at different levels of granularity.

[8]https://goo.gl/bc9dui

References

[1] M. Tsikerdekis, "Identity deception prevention using common contribution network data," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 1, pp. 188–199, Jan. 2017.

[2] T. Anwar and M. Abulaish, "Ranking radically influential Web forum users," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 6, pp. 1289–1298, Jun. 2015.

[3] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "Design and analysis of social botnet," *Comput. Netw.*, vol. 57, no. 2, pp. 556–578, 2013.

[4] D. Fletcher, "A brief history of spam," *TIME*, Nov. 2, 2009. [Online]. Available: http://www.time.com/time/business/article/0,8599,1933796,00.html

[5] Y. Boshmaf, M. Ripeanu, K. Beznosov, and E. Santos-Neto, "Thwarting fake OSN accounts by predicting their victims," in *Proc. AISec*, Denver, CO, USA, 2015, pp. 81–89.

[6] A. A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, and C. Yang, "CATS: Characterizing automation of Twitter spammers," in *Proc. COMSNETS*, Bengaluru, India, Jan. 2013, pp. 1–10.
[7] K. Lee, J. C. Lee, and S. Webb, "Uncovering social spammers: Social honeypots + machine learning," in *Proc. SIGIR*, Geneva, Switzerland, Jul. 2010, pp. 435–442.
[8] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proc. ACSAC*, Austin, TX, USA, 2010, pp. 1–9.
[9] H. Yu, M. Kaminsky, P. B. Gibbons, and A. D. Flaxman, "SybilGuard: Defending against sybil attacks via social networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 3, pp. 576–589, Jun. 2008.
[10] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *Proc. IMC*, Melbourne, VIC, Australia, 2001, pp. 35–47.
[11] W. Wei, F. Xu, C. C. Tan, and Q. Li "Sybildefender: Defend against sybil attacks in large social networks," in *Proc. INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 1951–1959.
[12] C. Yang, R. C. Harkreader, and G. Gu, "Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers," in *Proc. RAID*, Menlo Park, CA, USA, 2011, pp. 318–337.
[13] S. Lee and J. Kim, "WarningBird: A near real-time detection system for suspicious URLs in Twitter stream," *IEEE Trans. Depend. Sec. Comput.*, vol. 10, no. 3, pp. 183–195, May 2013.
[14] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," in *Proc. Workshop Learn. Text Categorization*, Madison, WI, USA, 1998, pp. 98–105.
[15] C. Schäfer, "Detection of compromised email accounts used by a spam botnet with country counting and theoretical geographical travelling speed extracted from metadata," in *Proc. ISSREW*, Naples, Italy, Nov. 2014, pp. 329–334.
[16] C. Schäfer, "Detection of compromised email accounts used for spamming in correlation with origin-destination delivery notification extracted from metadata," in *Proc. ISDFS*, Tirgu Mures, Romania, Apr. 2017, pp. 1–6.
[17] A. H. Wang, "Detecting spam bots in online social networking sites: A machine learning approach," in *Proc. DBSec*, Rome, Italy, 2010, pp. 335–342.
[18] F. Ahmed and M. Abulaish, "A generic statistical approach for spam detection in online social networks," *Comput. Commun.*, vol. 36, nos. 10–11, pp. 1120–1129, 2013.
[19] C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving Twitter spammers," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 8, pp. 1280–1293, Aug. 2013.
[20] Y. Zhu, X. Wang, E. Zhong, N. N. Liu, H. Li, and Q. Yang, "Discovering spammers in social networks," in *Proc. AAAI*, Toronto, ON, Canada, 2012, pp. 52–58.
[21] E. Tan, L. Guo, S. Chen, X. Zhang, and Y. Zhao, "Spammer behavior analysis and detection in user generated content on social networks," in *Proc. ICDCS*, Macau, China, Jun. 2012, pp. 305–314.
[22] S. Y. Bhat and M. Abulaish, "Community-based features for identifying spammers in online social networks," in *Proc. Int. Conf. Adv. Social Netw. Anal. Mining*, Niagara Falls, ON, Canada, Aug. 2013, pp. 100–107.
[23] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo, "People are strange when you're a stranger: Impact and influence of bots on social networks," in *Proc. AAAI*, Dublin, Ireland, 2012, pp. 10–17.
[24] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proc. WWW*, Perth, WA, Australia, 2017, pp. 963–972.
[25] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the Web," Stanford InfoLab, Stanford, CA, USA, Tech. Rep. 1999-66, Nov. 1999. [Online]. Available: http://ilpubs.stanford.edu:8090/422/
[26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.
[27] F. Benevenuto, G. Mango, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in *Proc. CEAS*, Redmond, WA, USA, 2010, pp. 55–65.
[28] C. J. Geyer, "Introduction to Markov chain Monte Carlo," in *Handbooks of Modern Statistical Methods*. London, U.K.: Chapman & Hall, 2011, pp. 1–46.
[29] M. Fire, G. Katz, and Y. Elovici, "Strangers intrusion detection-detecting spammers and fake profiles in social networks based on topology anomalies," in *Proc. ASE*, Essen, Germany, 2010, pp. 1–10.
[30] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, 2010.
[31] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Nat. Acad. Sci. USA*, vol. 103, no. 23, pp. 8577–8582, 2006.
[32] P. D. Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Mixing local and global information for community detection in large networks," *J. Comput. Syst. Sci.*, vol. 80, no. 1, pp. 72–87, 2014.
[33] C. Fawcett and H. H. Hoos, "Analysing differences between algorithm configurations through ablation," *J. Heuristics*, vol. 22, no. 4, pp. 431–458, 2016.
[34] S. C. Gupta and V. K. Kapoor, *Fundamentals of Methematical Statistics: A Modern Approach*. New Delhi, India: Sultan Chand & Sons, 2000.
[35] A. H. Wang, "Don't follow me: Spam detection in Twitter," in *Proc. SECRYPT*, Athens, Greece, Jul. 2010, pp. 1–10.

Mohd Fazil received the master's degree in computer science and application from Aligarh Muslim University, Aligarh, India, in 2013. He is currently pursuing the Ph.D. degree in computer science from Jamia Millia Islamia (A Central University), Delhi. He has qualified UGC-JRF exam in 2013. His research interests include data mining, data-driven cyber security, social network analysis, and machine learning. He is currently the recipient of the UGC Senior Research Fellowship.

Muhammad Abulaish (SM'12) received the Ph.D. degree in computer science from IIT Delhi in 2007. He is currently an Associate Professor and a Chairperson of the Department of Computer Science, South Asian University, Delhi. He has authored over 82 research papers in reputed journals and conference proceedings. His research interests span over the areas of data analytics and mining, social computing, and data-driven cyber security. He is a Senior Member of the ACM and CSI.