

# A Generic Statistical Approach for Spam Detection in Online Social Networks

Faraz Ahmed and Muhammad Abulaish<sup>1</sup>

Center of Excellence in Information Assurance  
King Saud University, Riyadh, Saudi Arabia  
{fAhmed.c, mAbulaish}@ksu.edu.sa

---

## Abstract

In this paper, we present a generic statistical approach to identify spam profiles on Online Social Networks (OSNs). Our study is based on real datasets containing both normal and spam profiles crawled from Facebook and Tweeter networks. We have identified a set of 14 generic statistical features to identify spam profiles. The identified features are common to both Facebook and Twitter networks. For classification task, we have used three different classification algorithms – *naïve Bayes*, *Jrip*, and *J48*, and evaluated them on both individual and combined datasets to establish the discriminative property of the identified features. The results obtained on a combined dataset has detection rate (DR) as 0.957 and false positive rate (FPR) as 0.048, whereas on Facebook dataset the DR and FPR values are 0.964 and 0.089, respectively, and that on Twitter dataset the DR and FPR values are 0.976 and 0.075, respectively. We have also analyzed the contribution of each individual feature towards the detection accuracy of spam profiles. Thereafter, we have considered 7 most discriminative features and proposed a clustering-based approach to identify *spam campaigns* on Facebook and Twitter networks.

*Keywords:* Data Mining, Social network analysis, Social network security, Spam profile identification, Spam campaign identification

---

<sup>1</sup>To whom correspondence should be made. Phone: +966-1-4696468, Fax: +96614695237, Email: abulaish@ieee.org

## 1. Introduction

Due to increasing popularity of social media, Online Social Networks (OSNs) have become a popular communication and information sharing tool over the past few years. The users of the social networks are the key role players and they are responsible for the contents being shared in the networks. The individual users are the basic elements in the hierarchy of the OSNs, and the next elements in the hierarchy of OSNs are the communities formed by friends, families, and acquaintances. Users share information by sharing links to interesting websites, videos, and files. Moreover, the community structure of OSNs creates a network of trust and reliability. An individual shares personal information with his/her network of trust and other users trust the information shared. A study done in [3] shows that 45% of users click on the links shared by their immediate contacts. This feature of sharing information to a large number of individuals with ease has attracted malicious parties, including *social spammers*. Social spammers exploit the network of trust for spreading spam messages promoting personal blogs, advertisements, phishing, and scam. Spammers employ different strategies for getting into a user's network of trust. Information sharing by the use of URL shortening service is an important feature of online social networking [4]. This feature is easily exploited and is particularly harmful to users if it contains links to scams advertisements, adult content, and other solicitations like phishing that attempts to capture account credentials, and pages attempting to distribute malware [21]. Therefore, the only security feature that protects a user from malicious parties is the network of trust.

According to [18], globally 75.9% of email messages are spam. Similarly, for the social networks the current state of spam is worsening and more rigorous efforts are required to stop them in an effective manner. Nowadays, spammers are trying a new approach to gain access through *Facebook events*. Generally, Facebook events are used by the spammers to invite users with bogus titles, e.g., “*check out who viewed your profile.*” Although, these links direct to valid Facebook event pages, once a user views more information, the malicious link is displayed<sup>2</sup>. Similarly, botnets, worms, and viruses have emerged on OSNs [9], [15]. The study of spam done in [17] and [25], point out different strategies used by bots to launch successful spam campaigns. Such spam campaigns consists of a single spammer having multiple accounts

---

<sup>2</sup><http://www.geeksugar.com/Spammers-Target-Facebook-Events-15447506>

on OSNs, which increases the chance of a user being exposed to spam [12].

In this paper, we present a generic approach to detect spam profiles on different categories of OSNs. Our study is based on real datasets collected from Facebook and Twitter networks that contain both benign and spam profiles. We have identified a set of 14 statistical features that are common to both Facebook and Twitter networks. The feature set consists of statistics related to wall posts/tweets, links shared, friends/followers, mentions, and hash tags that are calculated after logging the complete wall history of users' profiles. These features are the key elements to facilitates interactions in social networks, e.g., wall posts/tweets are the main form of communication between OSN users, similarly sharing URLs with friends/followers is the key source of information sharing. For identification of spam profiles, we have trained three different classification models – naïve Bayes, Jrip, and J48 on both combined and individual datasets crawled from Facebook and Twitter networks.

We have performed two different types of experiments to analyze the efficacy of the proposed method. In first experiment, the complete feature set is used to train the classifiers and test their classification accuracy using 10-fold cross validation. On separate datasets, the highest detection accuracy is 96.4% and 98.7% for Facebook and Twitter profiles, respectively, whereas the detection accuracy for the combined dataset containing spam-profiles of both Facebook and Twitter is 95.7%. In second experiment, the discriminative property of each individual feature is analyzed. For this, a feature is excluded from the feature set and classification algorithms are applied on remaining features to observe the increase/decrease in detection accuracy. This process is repeated for every feature. On the basis of discriminative property, 7 most discriminative features are selected to identify multiple spam profiles constituting a spam campaign. As a result, 3 and 4 different types of campaigns are identified in Twitter and Facebook datasets, respectively..

The rest of the paper is organized as follows. Section 2 presents an overview of the related works on social spam analysis. Section 3 describes our dataset collection methodology and the statistics of the dataset. Section 4 presents an explanation of identified features. Section 5 provides experimental setup and feature analysis, followed by spam campaign analysis in section 6. Finally, section 7 concludes the paper with some insights on the future directions of work.

## 2. Related Work

The huge amount of information residing on online social networking sites has attracted researchers to mine this information and study issues faced by the social network community. Considerable work has been done for collecting and mining information for various problems such as community detection, information diffusion, and spam filtering. In [16], the authors investigated the feasibility of using measurement calibrated graph models for sharing information among researchers without revealing private data. In [10], the authors presented a study of topological characteristics of Twitter network in which they investigated the behavior of information diffusion over the network by analyzing “retweets” and found that an information retweeted once reaches on average 1000 users. In [2], the authors presented a study on click-stream data of social networks and showed that the use of click-stream data provides rich information about social interactions. They also showed that a majority of user activities on social networks consists of *browsing*. Similarly, in [7], the authors investigated social interactions of users on OSNs and proposed that a majority of the interactions on OSNs are latent in nature, whereas visible events occur less frequently.

A number of research efforts has been also diverted towards the detection and prevention of spam on OSNs. In [19], the authors proposed a real time URL-spam detection scheme (Monarch) for Twitter in which they logged browser activities while loading a page for an URL. In this respect, they monitored a multitude of details including *redirects*, *domains contacted while constructing a page*, *HTML content*, *pop-up windows*, *HTTP headers*, and *JavaScript plugin execution* to detect spam links. As compared to our work, the URL spam filter is designed specifically for Twitter social network. Moreover, the focus of “Monarch” is to identify individual spam URLs present in tweets, and does not addresses the problem of detecting increasing number of spam profiles on the Twitter social network. Another substantial work on detection of spam on OSNs is presented in [17]. In this work, the authors created honey-profiles representing different age, nationality, etc. Their study is based on data collected from profiles of several regions, including USA, Middle East, Europe, etc. They logged all types of requests, wall posts, status updates, and private messages on Facebook. Similarly, on MySpace, they recorded mood updates, wall posts and messages, whereas on Twitter, they logged tweets and direct messages. Based on these activities, they developed six features to distinguish spam profiles from normal profiles. Most of these

features are related to a profile’s friending activity. In this work, we propose two new categories of features: the features pertaining to the statistical information on a profile’s Facebook page-likes and Facebook tags. In addition we provide a thorough analysis on the robustness of our features. The authors in [13] also utilized the concept of *social honeypot* to lure content polluters on Twitter, and the harvested users are analyzed to identify a set of features for classification purpose. Their technique is evaluated on a dataset of Twitter spammers collected using the *@spam* mention provided by Twitter to flag spammers. The authors have proposed four different categories of features specifically for the Twitter social network. These four types of categories consider statistics related to *followers/following*, *URLs*, *@mentions*, and *tweet content*; whereas, we propose an extra category of feature, i.e., the Twitter hashtags for identification of spam profiles. In [5], the authors analyzed a large dataset of wall posts on Facebook user profiles for detection of spam accounts. They built wall posts similarity graph for detection of malicious wall posts. The authors also presented an analysis of the profiles, which generate these wall posts. Our detection scheme focuses on the identification of profiles rather than posts generated by the profiles. Similarly, in [24], the authors presented a thorough analysis of profile-based and content-based evasion tactics employed by Twitter spammers. The authors proposed a set of 24 features consisting of graph-based, neighbor-based, automation based and timing based features that are evaluated using different machine learning techniques. The authors have also formalized the robustness of the proposed feature set.

In [6], the authors presented a large-scale effort to characterize spams on Twitter. Using click-through data generated from spam URLs, the authors analyzed the success of Twitter spam for luring over 1.6 million users to visit spam webpages. They clustered spam URLs present in tweets to identify trends that can distinguish *spam*, *malware* and *phishing*. In [14] and [8], the authors proposed combination of content-based and user-based features for detection of spam profiles on Twitter. In order to evaluate the importance of these features, the collected dataset is fed into traditional classifiers. A study of monetary relationships of spammers is given in [20]. This paper presents an analysis on the behavior of Twitter spammers. Based on a large Twitter dataset, the authors identified monetary relationships of spammers with vendors seeking to distribute their URLs. The authors also analyzed major spam campaigns and their life spans.

It can be observed in Table 1 that most of the previous works are specific

Table 1: Features, their categories and past references

Feature	Category	Feature used in:	
		Facebook	Twitter
f1	Interaction	[17]	[24], [23], [13]
f2	Interaction	New	New
f3	Posts/Tweets	New	[24], [14], [23]
f4	Posts/Tweets	New	New
f5	Posts/Tweets	New	New
f6	Posts/Tweets	[17]	N/A
f7	Posts/Tweets	New	N/A
f8	Posts/Tweets	New	N/A
f9	URLs	[17]	[24], [13], [14], [11]
f10	URLs	New	[24], [13], [11]
f11	URLs	New	New
f12	Tags/@mention	New	[13], [14], [23]
f13	Tags/@mention	New	[11]
f14	Tags/@mention	New	[13], [11]

to a particular type of network. To the best of our knowledge, none of the works related to Facebook have attempted to identify features for profile-level spam detection. Rather, the works targeting Facebook spam, consider features that are specific to the content of wall posts. Though such features are useful for identification of spam posts, for detection and reduction of spams on Facebook it is imperative to detect spammers that are spreading spams by using multiple fake Facebook profiles. Detection of a spam profile require a different set of features that define its behavior. For Twitter, some researchers have targeted profile-level spam detection, but most of them have analyzed URLs or click-through data to identify spams. And, even in the case of feature-based spam profile detection for Twitter, only tweet features are considered in the earlier works. In contrast to the existing approaches, our study presents a novel set of generic features for the detection of spam profiles on both Facebook and Twitter networks. The proposed feature set can be used to identify spam-profiles and thereby spammers to control the spread of spam contents on these networks.

### 3. Data Collection

To develop a dataset for training and testing of classification systems, we have manually identified a set of normal and spam profiles from both Facebook and Twitter networks. Since our proposed feature identification approach is based on visible interactions of the OSN users, we logged information only about the social interactions of their profiles.

### 3.1. Facebook Data

Facebook is the most popular social network claiming more than 800 million active users. The popularity of Facebook can be attributed to its platform features including *wall posts*, *fan pages* and *tags* that make social interactions and information sharing more interactive. Further details about these features are presented in the following paragraphs.

- **Wall Posts:** The Facebook wall of a user is a place where her friends or other Facebook users (depending on the privacy settings) can interact by posting messages and useful links. Users can also like and comment on the wall posts. According to Facebook statistics published in September 2011, about 2 billion wall posts on Facebook are liked or commented in a single day.
- **Fan pages:** Facebook fan pages are designed for celebrities, and business organizations who are interested to share information to people outside their real social circle. Users can **like** certain pages to get latest updates about their interests. A single user has indirect connection to larger groups of users via 80 (on average) community pages, groups and events.
- **Tags:** Facebook tagging feature allows user to tag friends and pages in posts (analogous to Twitter @mention). Once tagged in a post the content being shared becomes visible on the subject's wall and hence affects information diffusion.

For Facebook profiles, we logged users' activities on their Facebook wall. Only information available for public view was collected and users with restricted view of their profiles were not considered. We logged activities related to friendship requests, wall posts, fan page likes, and links shared. We logged only the visible interactions of the profiles. We logged this information from a total of 320 Facebook profiles, including 165 manually identified spam profiles and 155 normal user profiles. Our logging methodology utilized the publicly available Java API "HTML Parser", for gathering the required information. Each type of activity on a user's Facebook wall has its own identifiers in the HTML source. For example, the HTML structure of a URL shared by a user is different from the URL posted by someone else on the user's wall. We identified such structural differences and utilized them to gather the required details. However, the current Facebook "Timeline

profiles” have new HTML structures and our scripts are limited to profiles containing the Facebook “Wall”. A profile is categorized as spam on the basis of its visible activities. Spammers exhibit major wall post activity consisting of links directing to mostly fake pornographic websites, personal blogs, advertisements and so on. Another category of spam accounts we observed, consists of compromised accounts, infected or hacked accounts by malicious Facebook applications. Such accounts exhibit a plethora of posts sharing the same link directing to an advertisement campaign. Our spam dataset has a majority of spammers that use the Facebook tagging feature, and in each post Facebook friends and fan pages are tagged which make the link visible to more people than originally tagged.

### 3.2. Twitter Data

Twitter is one of the popular online social network, which supports micro-blogging – one of the main factors for Twitter popularity. A Twitter user can use **tweets**, **mentions** or **hashtags** for information sharing with her *followers*.

- **Tweets:** This feature facilitates Twitter users to disseminate information by sharing a link or an update written using maximum 140 characters.
- **@mentions:** This feature allows Twitter users to directly address someone by using the *@username* methodology.
- **Hash-tags:** The hash-tags feature allows Twitter users to keep track of tweets that are related to certain keywords. Hence users can get updates of their interest by looking for tweets via hashtags.

For Twitter, we logged similar information from a total of 305 profiles, including 160 spam profiles and 145 normal user profiles. The data logged consists of all tweets of a user since the creation of her profile. We used the same Java API “HTML Parser” for gathering details from Twitter profiles. For each profile we gathered names of hash-tags used, @mentions and links tweeted. As we focus on actual interactions of a user, we have gathered details of other users *@mentioned* by her, and we do not consider complete list of following/followers. Twitter spammers exhibit strongly similar behavior such as same links tweeted several times, same hash-tags and mentions. Statistics related to the information collected is given in Table 2. Facebook



Table 2: Statistics of Facebook and Twitter profiles

Category	Links	Likes/Hashtags	Friends/Mentioned
Facebook normal	20175	21975	42124
Facebook spam	53836	67536	107953
Twitter normal	24373	32950	29925
Twitter spam	83076	395	494

spam profiles show significantly greater activity as compared to Twitter spam profiles. Twitter spammers use very few hash-tags usually the popular ones for spreading spam. However tweeting links is the most dominant activity of Twitter spammers.

#### 4. Statistical Features

In order to model a classification system to classify Facebook and Twitter profiles as spam or normal, we have analyzed the crawled data to identify a common set of discriminative statistical features for both type of networks. We generalize social networks as a combination of 4 basic components – interactions, posts/tweets, URLs, and tags/mentions, and the activity of a user can be assigned to any one of these sectors. In the following sub-sections, we present a detail description of these components and the derived features for both Facebook and Twitter networks.

##### 4.1. Facebook Features

For Facebook, we have identified a total of 14 features (numbered as  $f_1, f_2, \dots, f_{14}$ ) that quantify the activity of a profile in each of the 4 sectors. A brief description of each feature is given in the following sub-sections.

##### 4.1.1. Interactions-Driven Facebook Features

When we focus on visible social interactions of a profile, we identify two entities to which users can interact on the Facebook social network – *friends* and *community pages*, which are identified as main features in this sector.

**Friends ( $f_1$ ):** For each profile, we calculate the number of friends who have been visibly interacted by or who have interacted with the subject profile. For example, when a user  $u_i$  posts something on a friend’s Facebook wall, this interaction is visible on  $u_i$ ’s wall in the form of a small message and the main content is visible on the friend’s wall. Such interaction is a user-friend interaction. Similarly, if a friend posts on  $u_i$ ’s wall then the complete content is visible and the friend is identified as an active friend. This is a

friend-user interaction. Other kinds of interaction activity visible on a user's wall include tags, comments, and likes. *Tags* and *comments* are also two-way interaction features. Interaction between a user and its friends through these features is visible on the subject profile's wall. *Like* feature is a user-friend interaction method in which a friend *Likes* a post or a comment already present on the user's wall.

**Community pages ( $f_2$ ):** For each profile, we calculate the number of community pages in which the user has been actively participating. In case of Facebook community pages, only user-page interaction is possible. Users have to *like* a page to participate in the page's activity. Users can participate through posts, comments, tags, and likes. When a user posts on a *page's* wall, the activity is visible on the user's wall in the form of a message.

#### 4.1.2. Posts-Driven Facebook Features

The major part of communication on Facebook is done through posts. We categorize wall posts as *page-posts* and *profile-posts*. Page-posts correspond to a user's posting activity on a community page's wall and profile-posts are the user's posts on its friends' walls. We have identified six features to analyze wall post activities.

**Page-posts:** Popular pages have huge number of participants and they are more likely to be targeted by spammers and it can be hard for the page administrators to manually identify and report the spammers. Based on the page-post activity of individual users, we have identified three features ( $f_3$  through  $f_5$ ) that can provide important information for classifying user profiles.

- $f_3$ : This feature represents the total number of posts generated by a user on her community pages.
- $f_4$ : This feature represents the maximum number of posts by a user on her community pages.
- $f_5$ : This represents the rate of posts, i.e., number of posts per page.

**Profile-posts:** Based on profile-posts, we have identified following three features ( $f_6$  through  $f_8$ ):

- $f_6$ : This feature represents the total number of posts generated by a user on its friends' walls.

- $f_7$ : This feature is used to represent the maximum number of posts by a user generated on her friends' walls.
- $f_8$ : This feature represents the rate of posts, i.e., number of posts per friend.

#### 4.1.3. URL-Driven Facebook Features

URLs of different websites are the main source of information on online social networks. Facebook users *share* website URLs with their friends for spreading interesting information. On a Facebook user's wall, URLs can be found in posts and comments. However, spammers have used this URL sharing feature to share URLs, which direct users to malicious websites, personal blogs, and so on. In our analysis, we log all un-trusted website links that are shared by a user. All links published by Facebook and its advertisements are considered trusted. We do not consider URLs that are shared by a user's friends on her wall. This is because, if a user has a lot of spam posts containing URLs that are shared by a friend then considering such URL information can classify the user's profile as spam. The following three features ( $f_9$  through  $f_{11}$ ) are related to URLs:

- $f_9$ : This feature represents the total number of URLs shared by a user.
- $f_{10}$ : This represents the number of unique URLs. For each profile, we identify the number of URLs that have been shared at least once by the user. This feature, together with the total number of URLs shared, model the URLs sharing behavior of a user.
- $f_{11}$ : This feature captures the average URL repetition frequency. For this, we generate the frequency histogram of each unique URL and the average frequency depicts the user behavior towards URL sharing.

#### 4.1.4. Tags-Driven Facebook Features

Facebook users can *tag* friends and pages. Tagging a friend has effects similar to a profile-post and tagging a page is similar to a page-post. This feature simplifies the method of information sharing, i.e., the post can be shared by just typing a friend's name preceded by the symbol @. Spammers tag multiple users or pages in a single post to spread the content to a larger community with less effort. For each profile, we calculate the following three features ( $f_{12}$  through  $f_{14}$ ) related to Facebook tags:

- $f_{12}$ : This represents the total number of tags present in the posts shared by a user.
- $f_{13}$ : This represents the total number of users and pages tagged.
- $f_{14}$ : This feature represents the rate of tagging, i.e., average number of tags present in a post.

#### 4.2. Twitter Features

In this section, we have tried to map the previously identified 14 Facebook features to Twitter social network and analyze their discriminative properties, so that the same set of features could be used to characterize both Facebook and Twitter profiles. Although, due to being different networks, Facebook and Twitter provides different terminologies to characterize users activities, we have used the same feature numbering scheme ( $f_1, f_2, \dots, f_{14}$ ) to show their equivalence in both the networks. In line with the identification process of Facebook features, we have considered *interactions*, *tweets*, and *URLs* to derive Twitter features, whose further details are presented in the following sub-sections.

##### 4.2.1. Interactions-Driven Twitter Features

Since, interaction between Twitter users is based on entities similar to that of Facebook users, we map the Facebook *friend* and *community page* features to *followers* and *hash tags*, respectively.

**Followers** ( $f_1$ ): For each profile, this feature represents the number of followers. As Twitter is a simpler platform than the Facebook, there are fewer ways through which visible interactions can be identified. Therefore, we consider the actual number of followers mentioned on a user’s profile. As compared to *friends* in Facebook, *followers* of a user can also receive tweets (updates) from other users and can mention (post) the users in their tweets.

**Hash-tags** ( $f_2$ ): This feature is considered to capture the total number of unique hash-tags present in a user’s profile. Hash-tags are similar to *page-likes* as users interested in a particular hash-tag can search for tweets over the Twitter network containing the particular hash-tag. Moreover, a user can make its tweet visible to a larger community by using a hash-tag which is similar to the *page-post* feature of Facebook.

#### 4.2.2. Tweets-Driven Twitter Features

The basic form of communication in the Twitter network is the *tweeting* feature. Users can share information by using features such as *hash-tags* and *@mentions*. Consequently, we have identified three features from each for characterizing users tweeting activities.

**Hash-tags:** We map the three *page-post* features of Facebook to hash-tags. Each tweet containing a hash-tag can be considered as a page-post, and consequently the features  $f_3$ ,  $f_4$ , and  $f_5$  for Twitter network users can be defined as follows:

- $f_3$ : This is defined as the total number of hash-tags used by a user.
- $f_4$ : This represents the maximum value of the frequency histogram generated for hash-tags used in a profile.
- $f_5$ : This is similar to the posting rate feature of Facebook, and it is defined to represent the hash-tagging rate of a profile. This is calculated as the average of the frequency histogram of the hash-tags used in a profile.

**@mentions:** For each profile, we calculate three features related to @mentions. Facebook *profile-post* features are similar to @mentions, but when compared to the Facebook's tagging feature, @mentions and tags have same functionalities. For example, like *tags* in Facebook which allows information sharing with multiple users, @mentions can be used to post a single tweet on Twitter-timeline of multiple users. Thus, each @mentions feature, given below, is a mapping of *tags* feature of Facebook profiles.

- $f_{12}$ : This is used to represent the total number of times a user has used @mention in her tweets.
- $f_{13}$ : For this, we generate a frequency histogram of @mentions used in a profile. A combination of information from this feature and the rate of using @mentions can help in learning the @mentions usage behavior of a user.
- $f_{14}$ : This represents the rate of @mentioning, i.e., number of @mentions per friend. This feature also exhibits coherence with  $f_{14}$  feature of Facebook.

In case of Twitter, the features  $f_6$ ,  $f_7$ , and  $f_8$  are not applicable and their values can be considered as nulls. Though these profile-posts-based features are important for thoroughly analyzing the behavior of Facebook profiles, due to simplicity of the Twitter network they cannot be logically mapped to Twitter profiles. In section 5, we present a thorough analysis to highlight the contribution of each feature for detecting spam profiles.

#### 4.2.3. URL-Driven Twitter Features

On Twitter network, users can use maximum 140 characters to share information. Due to this limitation, users generally utilize URL shortening services to tweet URLs. Twitter spammers use this feature to tweet URLs, which direct users to malicious phishing websites. Like Facebook features  $f_9$ ,  $f_{10}$ , and  $f_{11}$ , we have defined three statistical features derived from URLs as follows:

- $f_9$ : This represents the total number of URLs shared by a user.
- $f_{10}$ : This is defined as the number of unique URLs.
- $f_{11}$ : This is similar to the Facebook’s  $f_{11}$  feature in which we calculate the average URL repetition frequency.

### 5. Experiment Setup and Feature Evaluation Results

In this section, we present a thorough evaluation of the identified features on both individual and combined datasets of Facebook and Twitter. We have considered three different classification algorithms – *naïve Bayes (NB)*, *rule learner (Jrip)*, and *decision tree (J48)*, to establish the discriminative properties of the identified features to classify spam and benign profiles. Table 3 shows the Detection Rate ( $DR$ ) and False Positive Rate ( $FPR$ ) achieved after applying these classification algorithms. Naïve Bayes gives the best results for Facebook dataset, Jrip proves best for Twitter dataset, and the tree based algorithm J48 shows best results for the combined dataset. In order to further analyze the contribution of each feature, we utilize Information Gain (IG) to quantify the importance of each feature. Table 4 gives the IG values of each feature for every category of the dataset. We also perform another set of experiments to further analyze the contribution of each feature towards Facebook and Twitter datasets.

Table 3: Performance evaluation results of different classifiers on individual and combined datasets

Algorithms	NB		Jrip		J48	
	FPR	DR	FPR	DR	FPR	DR
<b>Facebook</b>	0.089	<b>0.964</b>	0.09	0.912	0.081	0.898
<b>Twitter</b>	0.075	0.976	0.014	<b>0.987</b>	0.017	0.983
<b>Combined</b>	0.309	0.733	0.071	0.935	0.048	<b>0.957</b>

Table 4: Information gain values of features

Feature	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14
<b>FB</b>	0.15	<b>0.33</b>	0.17	0.09	0.14	0.09	0.09	0.13	0	0	0.31	0.18	0.21	0.25
<b>T</b>	0.32	0.82	<b>0.87</b>	0	0	0	0	0.11	0.17	0.19	0.40	0.86	0.83	0.57
<b>FB+T</b>	0.11	0.07	0.11	0.11	0.17	0.06	0.04	0.04	0.08	0.05	0.13	0.08	0.10	<b>0.24</b>

In the next stage of analysis, we have removed one feature at a time from the feature set (F) and used the resultant feature set to apply classification algorithms. Table 5, 6 and 7 shows the detection rates achieved for naïve Bayes, Jrip and J48 classification algorithms. In these tables, the values in parenthesis show the amount of increase/decrease in false positive rates (FPR) and detection rates (DR) after removing the corresponding feature from the feature set. A brief discussion about the evaluation results obtained for each classification algorithm is presented in the following sub-sections.

### 5.1. Naïve Bayes Classification

Naïve Bayes being a simple probabilistic classifier, treats every feature independently. In other words, it assumes that a particular feature is independent of the values of any other feature. Table 5 shows the FPR and the DR achieved through applying naïve Bayes algorithm for each feature subset on different datasets. It can be seen in this table that after removing a feature from the feature set there is an increase or decrease in FPR and DR values according to the importance of the removed feature. Further observations related to the independence of features are enumerated in the following paragraphs:

- The results obtained after the removal of feature  $f_1$  shows that FPR is reduced in case of individual dataset, hence treating  $f_1$  as an independent feature is not beneficial. However, on combined dataset, it increases the false positives.
- Feature  $f_3$  (number of page posts) is obviously very important for Facebook dataset, because its removal significantly increases FPR and reduces DR. This shows that number of page-posts, significantly affects

Table 5: Feature evaluation using naïve Bayes algorithm

Dataset	Facebook		Twitter		Combined	
Parameters	FPR	DR	FPR	DR	FPR	DR
$F - f_1$	0.080(↓ 0.009)	0.961(↓ 0.003)	0.048(↓ 0.027)	0.976(0.000)	0.326(↑ 0.017)	0.741(↑ 0.008)
$F - f_2$	0.083(↓ 0.006)	0.963(↓ 0.001)	0.075(0.000)	0.977(↑ 0.001)	0.303(↓ 0.006)	0.741(↑ 0.008)
$F - f_3$	0.196(↑ 0.107)	0.881(↓ 0.083)	0.072(↓ 0.003)	0.980(↑ 0.004)	0.300(↓ 0.009)	0.746(↑ 0.013)
$F - f_4$	0.080(↓ 0.009)	0.964(0.000)	0.075(0.000)	0.976(0.000)	0.303(↓ 0.006)	0.740(↑ 0.003)
$F - f_5$	0.080(↓ 0.009)	0.965(↑ 0.001)	0.09(↑ 0.015)	0.983(↑ 0.007)	0.320(↑ 0.011)	0.725(↓ 0.008)
$F - f_6$	0.086(↓ 0.003)	0.962(↓ 0.002)	0.075(0.000)	0.976(0.000)	0.306(↓ 0.003)	0.737(↑ 0.004)
$F - f_7$	0.092(↑ 0.003)	0.968(↑ 0.004)	0.075(0.000)	0.976(0.000)	0.309(0.000)	0.733(0.000)
$F - f_8$	0.092(↑ 0.003)	0.962(↓ 0.002)	0.075(0.000)	0.976(0.000)	0.306(↓ 0.003)	0.728(↓ 0.005)
$F - f_9$	0.080(↓ 0.009)	0.964(0.000)	0.092(↑ 0.017)	0.978(↑ 0.002)	0.316(↑ 0.007)	0.730(↓ 0.003)
$F - f_{10}$	0.098(↑ 0.009)	0.964(0.000)	0.089(↑ 0.014)	0.977(↑ 0.001)	0.326(↑ 0.017)	0.722(↓ 0.011)
$F - f_{11}$	0.086(↓ 0.003)	0.953(↓ 0.011)	0.090(↑ 0.015)	0.976(0.000)	0.321(↑ 0.013)	0.723(↓ 0.010)
$F - f_{12}$	0.080(↓ 0.009)	0.964(0.000)	0.067(↓ 0.008)	0.967(↓ 0.009)	0.294(↓ 0.015)	0.749(↑ 0.016)
$F - f_{13}$	0.080(↓ 0.009)	0.963(↓ 0.001)	0.108(↑ 0.033)	0.977(↑ 0.001)	0.293(↓ 0.016)	0.749(↑ 0.016)
$F - f_{14}$	0.095(↑ 0.006)	0.963(↓ 0.001)	0.097(↑ 0.022)	0.971(↓ 0.005)	0.325(↑ 0.016)	0.701(↓ 0.032)

the decision of naïve Bayes algorithm. However, in case of Twitter, there is a slight improvement in the results after removing the feature. Hence, as a singular feature, Twitter  $f_3$  does not reveal high classification power. Similarly, for combined dataset the results improve slightly after removal of  $f_3$ , which reveals its low strength for combined dataset.

- URL-driven features ( $f_9$ ,  $f_{10}$ , and  $f_{11}$ ) are important for Twitter dataset as there is an increase in FPR when any of these features is removed. Since URLs are the only source of information in the Twitter social network, the importance of URLs-related features cannot be neglected. As naïve Bayes assumes independence among features, we can say that each feature contributes independently in the classification process. Similar effects are reflected in case of combined dataset.
- @mention being another important Twitter feature, has prominent effects on classification of Twitter profiles. We can see that the absence of feature  $f_{13}$  or  $f_{14}$  causes an increase in FPR.

### 5.2. Jrip Classification

We chose the rule-based learner, Jrip, due to its inherent simplicity that results in a better understanding of the learned model. The algorithm works by initially developing a set of rules for decision making and improving the rules iteratively using different heuristic techniques. The final rule set is used to classify test cases. Figure 1 shows few sample rules generated by Jrip for each dataset category. Table 6 shows the feature evaluation result using Jrip



Table 6: Feature evaluation using Jrip algorithm

Dataset Parameters	Facebook		Twitter		Combined	
	FPR	DR	FPR	DR	FPR	DR
$F - f_1$	0.085(↓ 0.005)	0.919(↑ 0.007)	0.025(↑ 0.011)	0.980(↓ 0.007)	0.11(↑ 0.040)	0.889(↓ 0.046)
$F - f_2$	0.101(↑ 0.011)	0.898(↓ 0.014)	0.043(↑ 0.029)	0.957(↓ 0.030)	0.090(↑ 0.019)	0.918(↓ 0.017)
$F - f_3$	0.096(↑ 0.006)	0.910(↓ 0.002)	0.024(↑ 0.010)	0.978(↓ 0.009)	0.087(↑ 0.016)	0.928(↓ 0.007)
$F - f_4$	0.087(↓ 0.003)	0.912(0.000)	0.021(↑ 0.007)	0.981(↓ 0.006)	0.090(↑ 0.019)	0.925(↓ 0.010)
$F - f_5$	0.087(↓ 0.003)	0.908(↓ 0.004)	0.021(↑ 0.007)	0.979(↓ 0.008)	0.087(↑ 0.016)	0.923(↓ 0.012)
$F - f_6$	0.087(↓ 0.003)	0.915(↑ 0.003)	0.021(↑ 0.007)	0.979(↓ 0.008)	0.089(↑ 0.018)	0.937(↑ 0.002)
$F - f_7$	0.093(↑ 0.003)	0.916(↑ 0.004)	0.021(↑ 0.007)	0.979(↓ 0.008)	0.090(↑ 0.019)	0.928(↓ 0.007)
$F - f_8$	0.107(↑ 0.017)	0.912(0.000)	0.014(0.000)	0.984(↓ 0.003)	0.083(↑ 0.012)	0.920(↓ 0.015)
$F - f_9$	0.087(↓ 0.003)	0.914(↑ 0.002)	0.024(↑ 0.010)	0.980(↓ 0.007)	0.075(↑ 0.004)	0.925(↓ 0.010)
$F - f_{10}$	0.079(↓ 0.011)	0.908(↓ 0.004)	0.014(0.000)	0.990(↑ 0.003)	0.094(↑ 0.023)	0.922(↓ 0.013)
$F - f_{11}$	0.112(↑ 0.022)	0.872(↓ 0.040)	0.021(↑ 0.007)	0.979(↓ 0.008)	0.072(↑ 0.001)	0.938(↑ 0.003)
$F - f_{12}$	0.077(↓ 0.013)	0.923(↑ 0.011)	0.017(↑ 0.003)	0.983(↓ 0.004)	0.084(↑ 0.013)	0.930(↓ 0.005)
$F - f_{13}$	0.08(↓ 0.010)	0.923(↑ 0.011)	0.024(↑ 0.010)	0.975(↓ 0.012)	0.087(↑ 0.016)	0.913(↓ 0.022)
$F - f_{14}$	.077(↓ 0.013)	0.917(↑ 0.005)	0.017(↑ 0.003)	0.987(0.000)	0.090(↑ 0.019)	0.928(↓ 0.007)

algorithm. The results obtained are similar to the results of naïve Bayes. For Jrip, the features related to URLs and @mentions are important as there is an increase in FPR value when these features are removed.

#### Twitter

```
(f3 >= 34)&&(f2 >= 27) => class=norm (138.0/0.0)
```

```
(f3 >= 3)&&(f9 <= 77) => class=norm (4.0/0.0)
```

```
=> class=spam (160.0/1.0)
```

#### Facebook

```
(f3 <= 31)&&(f13 <= 31) and(f11 <= 1) => class=norm (147.0/6.0)
```

```
(f2 <= 57)&&(f3 <= 100)&&(f1 >= 4) => class=norm (11.0/3.0)
```

```
=> class=spam (162.0/5.0)
```

#### Combined

```
(f3 >= 1)&&(f13 >= 4)&&(f14 <= 1) => class=norm (164.0/6.0)
```

```
(f8 >= 1)&&(f3 <= 46)&&(f6 <= 18) => class=norm (61.0/1.0)
```

```
(f2 <= 81)&&(f9 <= 81)&&(f3 <= 36)&&(f6 <= 0)&&(f11 <= 2) => class=norm (55.0/1.0)
```

```
(f10 >= 80)&&(f10 <= 309)&&(f2 <= 133)&&(f3 <= 10) => class=norm (10.0/1.0)
```

```
=> class=spam (332.0/16.0)
```

Figure 1: Jrip Rules

### 5.3. Decision Tree (J48) Classification

The decision tree algorithm, J48, performs classification by creating a decision tree based on the features of the input training data. The root node of the tree is the feature with the highest information gain, i.e., it has the maximum classification power. The leaf node describes the decision

<pre> f3 &lt;= 46   f13 &lt;= 51     f11 &lt;= 1: norm (154.0/9.0)     f11 &gt; 1       f11 &lt;= 4         f3 &lt;= 10: norm (7.0/1.0)         f3 &gt; 10: spam (3.0)       f11 &gt; 4: spam (7.0)   f13 &gt; 51: spam (48.0) f3 &gt; 46: spam (101.0/3.0) </pre>	<pre> f13 &lt;= 7   f3 &lt;= 4: spam (147.0)   f3 &gt; 4     f9 &lt;= 210: norm (2.0)     f9 &gt; 210: spam (6.0) f13 &gt; 7   f3 &lt;= 12     f9 &lt;= 83: norm (3.0)     f9 &gt; 83: spam (6.0)   f3 &gt; 12: norm (138.0) </pre>
--	---

(a) Decision tree generated from Facebook dataset (b) Decision tree generated from Twitter dataset

Figure 2: Sample decision trees generated by J48 for Facebook and Twitter datasets

of the algorithm. Hence, the value of the leaf node is dependent on other independent nodes of the tree. Sample decision trees generated for each dataset are shown in figures 2(a), 2(b) and 3. Table 7 shows the feature evaluation results obtained through applying J48 classification algorithm. Some of our important observations are enumerated below.

- In case of combined dataset, all the features contribute to some extent in decision making process. Figure 3 shows some important nodes of the decision tree generated from the combined dataset. In can be seen that every feature has some contribution in the generation of the decision tree.
- The results obtained after removing feature  $f_1$  show that the contribution of the feature towards separate Facebook and Twitter datasets is not significant. However,  $f_1$  proves to be an important node in the decision tree formed for the classification of combined dataset. This observation is coherent with the results obtained through naïve Bayes algorithm.
- Removal of feature  $f_3$  shows that it is important for the classification of all categories of dataset. In case of Facebook, naïve Bayes shows similar results. Moreover, the decision tree created from the Twitter dataset highlights the importance of  $f_3$  as a node. Figure 2 shows that for both



Table 7: Feature evaluation using J48 decision tree algorithm

Dataset Parameters	Facebook		Twitter		Combined	
	FPR	DR	FPR	DR	FPR	DR
$F - f_1$	0.081(0.000)	0.898(0.000)	0.017(0.000)	0.983(0.000)	0.091( $\uparrow$ 0.043)	0.933( $\downarrow$ 0.024)
$F - f_2$	0.078( $\downarrow$ 0.003)	0.916( $\uparrow$ 0.018)	0.017(0.000)	0.983(0.000)	0.079( $\uparrow$ 0.031)	0.927( $\downarrow$ 0.030)
$F - f_3$	0.081(0.000)	0.904( $\uparrow$ 0.008)	0.017(0.000)	0.983(0.000)	0.089( $\uparrow$ 0.041)	0.942( $\downarrow$ 0.015)
$F - f_4$	0.094( $\uparrow$ 0.013)	0.919( $\uparrow$ 0.021)	0.044( $\uparrow$ 0.027)	0.957( $\downarrow$ 0.026)	0.081( $\uparrow$ 0.033)	0.932( $\downarrow$ 0.025)
$F - f_5$	0.081(0.000)	0.898(0.000)	0.017(0.000)	0.983(0.000)	0.081( $\uparrow$ 0.033)	0.927( $\downarrow$ 0.030)
$F - f_6$	0.081(0.000)	0.898(0.000)	0.017(0.000)	0.983(0.000)	0.079( $\uparrow$ 0.031)	0.934( $\downarrow$ 0.023)
$F - f_7$	0.084( $\uparrow$ 0.003)	0.893( $\downarrow$ 0.005)	0.017(0.000)	0.983(0.000)	0.080( $\uparrow$ 0.032)	0.934( $\downarrow$ 0.023)
$F - f_8$	0.081(0.000)	0.898(0.000)	0.017(0.000)	0.983(0.000)	0.079( $\uparrow$ 0.031)	0.933( $\downarrow$ 0.024)
$F - f_9$	0.081(0.000)	0.898(0.000)	0.034( $\uparrow$ 0.017)	0.971( $\downarrow$ 0.012)	0.077( $\uparrow$ 0.029)	0.936( $\downarrow$ 0.021)
$F - f_{10}$	0.081(0.000)	0.898(0.000)	0.017(0.000)	0.983(0.000)	0.075( $\uparrow$ 0.027)	0.924( $\downarrow$ 0.033)
$F - f_{11}$	0.091( $\uparrow$ 0.010)	0.914( $\uparrow$ 0.016)	0.017(0.000)	0.983(0.000)	0.083( $\uparrow$ 0.035)	0.926( $\downarrow$ 0.031)
$F - f_{12}$	0.081(0.000)	0.898(0.000)	0.017(0.000)	0.983(0.000)	0.079( $\uparrow$ 0.031)	0.933( $\downarrow$ 0.024)
$F - f_{13}$	0.081(0.000)	0.903( $\uparrow$ 0.005)	0.027( $\uparrow$ 0.010)	0.975( $\downarrow$ 0.008)	0.079( $\uparrow$ 0.031)	0.929( $\downarrow$ 0.028)
$F - f_{14}$	0.071( $\downarrow$ 0.010)	0.913( $\uparrow$ 0.015)	0.021( $\uparrow$ 0.004)	0.980( $\downarrow$ 0.003)	0.092( $\uparrow$ 0.044)	0.938( $\downarrow$ 0.019)

#### 5.4. Discussions

In this section, we discuss the observations related to the importance of every feature for Facebook and Twitter social networks.

$f_1$ : According to our observations, normal users interact with a small portion of their Facebook friends, usually the ones who are more active or have similar interests. Moreover, a normal user’s wall exhibit two-way interactions. However, some spammers exhibit mostly one way interactions with majority of their friends. Spammers add a large number of users and spam them through various interaction related features.

$f_2$ : We observed that normal users either join a limited number of Facebook pages or show minimal activity, which can reveal their participation. However, spammers join large number of community pages and exhibit notable activity on most of the pages. To spread spam content more effectively, spammers *like* large number of pages and post spam contents. Consequently, the malicious post becomes visible to all members of the page. Therefore, through just monitoring a profile’s activity on Facebook pages, malicious behavior can be identified to some extent.

$f_3$ : Normal users generate limited Facebook page posting activity; however, spammers target popular community pages and post more as compared to normal users. A larger number of posts indicates that a profile is actively involved in community pages. Hence, page-post feature can be useful in distinguishing spammers from normal users. In case of Twitter, we consider hash-tags analogous to Facebook page. Normal users use a variety of hash-tags, whereas spammers use only popular hash-tags and the total number of

hash-tags is usually greater than that of the normal users.

$f_4$ : This feature along with the information about post rate helps to identify Facebook profile behavior. For example, a profile exhibiting large value with high post rate depicts a spam behavior. In case of Twitter, this feature along with the information about hash-tagging rate helps in identifying the spamming behavior of a profile. For example, a profile exhibiting large maximum value with higher hash-tagging rate depicts that same hash-tag is being used extensively and can be considered as a spamming behavior.

$f_5$ : A high post rate with a large number of Facebook page-likes depicts a spam behavior. Normal users do not tend to *like* a large number of pages. Hence, having a high post rate together with a large number of page-likes depicts spam behavior. Similarly for Twitter a high rate of hash-tagging with a small number of hash-tags depicts a spamming behavior.

$f_6$ : On analysis, we found that normal users on Facebook do not generate large number of posts as compared to spammers. This is because normal users communicate with a small group of active friends and also employ other methods of interaction such as comments and private messages. Whereas, spammers mostly utilize a single and effective method of posting malicious contents on their friends' walls. It is effective because a wall post becomes visible to the friends of the target profile, and increases the number of users exposed to spam.

$f_7$ : This feature is analogous to the *page-post* feature as it provides important insights about a profile's behavior by utilizing information about posting rate of the profile.

$f_8$ : Spammers exhibit extreme values of this feature, i.e., they have either a higher posting rate with a large number of friends (satisfying the conditions explained in section 4.1.1), or they have very low posting rate which shows that an entirely different spamming strategy is being used. For example, some spammers *tag* users and pages in spam posts.

$f_9$ : Our analysis of spam and normal Facebook profiles reveals that normal users share small number of untrusted URLs, including links to mostly popular video sharing sites, whereas spammers share large number of URLs, usually directing to a few different websites. Twitter spammers share large number of URLs, generally directing to the same website. This behavior is strongly similar to the Facebook spammers.

$f_{10}$ : Small number of unique URLs with a high number of sharing indicates a spamming activity on Facebook. Similarly, for Twitter, this feature together with the total number of URLs shared models Twitter spammers.

For example, a small number of unique URLs with a large number of URLs shared indicates an spamming activity.

$f_{11}$ : A higher value of this feature means that the same URL has been shared multiple times and hence is a probable spamming behavior.

$f_{12}$ : Normal users *tag* their friends and Facebook pages in posts and comments. However, the number of tags present in a normal user's profile is less as compared to that of a spammer's profile. Spammers, exploiting the tagging feature, tag a large number of users and pages. On Twitter, we observed that normal users do not use large number of *@mentions* as compared to spammers. This is because normal users communicate with a small group of followers/following users who are active on Twitter, whereas spammers mostly *@mention* every Twitter contact regardless of a user's activity status.

$f_{13}$ : This feature reveals the extent to which the tagging feature has been used by a Facebook profile. A large value indicates that the content is directly shared with a large community of people. Similarly, for Twitter a large value indicates that the user is actively involved in a conversation and can be considered normal. However, for well-established spammers this value can confuse the above explanation. In this case, the rate with which other users are @mentioned can clear the disambiguations.

$f_{14}$ : A higher rate depicts spamming behavior because spammers exploiting the tagging feature of Facebook try to tag multiple users and pages in a single post. Twitter spam profiles have extreme values for this feature, i.e., they have either a higher mentioning rate with a large number of followers/following or they have very small values, which shows that an entirely different spamming strategy is being used. For example, some spammers use hash-tags for effective spamming. Moreover, the low @mentioning rate can save the spammers from being reported. When a user is @mentioned in a spam tweet directly from one of its followers, it is more likely that the user will report the spam profile.

## 6. Spam Campaign Analysis

Our analysis on individual features shows that features related to friends/followers, pages/Hash-tags and URLs are important for classification in most of the cases. Removing such a feature from the dataset results in reduced detection accuracy, which is due to misclassification of certain instances. Our analysis also reveals that some features provide critical information about profiles' behavior. Spammers generating a spam campaign, generally make

use of multiple profiles and exploit various social networking features. For example, a group of spam profiles sharing a single malicious link multiple times. Such a behavior is easily identifiable by analyzing the feature values.

In order to identify spam campaigns, we model profiles using a weighted-graph in which spam profiles are inter-linked based on the information contained in the identified set of features. We utilize the modeling approach employed in one of our previous works [1]. We model an undirected social graph  $G = (V, E)$ , in which each node  $v_i$  represents a spam profile. An edge of the graph represents the connection between two spam profiles. Given a set of nodes  $V = \{v_i : i = 1, \dots, n\}$  and the edges  $E = \{(v_i, v_j) : 1 \leq i, j \leq n\}$ , let  $A$  be a  $n \times n$  similarity matrix representing nodes similarity in the graph. In this matrix,  $A(i, j) = a_{ij} \geq 0$  represents the similarity value of the corresponding nodes  $v_i$  and  $v_j$ , i.e.,  $a_{ij}$  represents the weight of the edge  $v_i$  and  $v_j$ . Each edge is represented by a vector consisting of values related to three categories of features – friends/followers ( $f_1$ ), page/hash-tags ( $f_3, f_4, f_5$ ) and URLs ( $f_9, f_{10}, f_{11}$ ).

We have applied Markov clustering algorithm [22] on the generated social graph to group spam profiles that exhibit similar activities. Markov clustering method uses a random walk on a weighted graph. It calculates the probability of moving from one node to another in an undirected graph. The probability of intra-cluster transitions is greater than inter-cluster transitions. So for a similarity matrix  $A$  of the graph  $G = (V, E)$ , the normalized adjacency matrix  $M$  is the transition matrix for a Markov random walk and  $M(i, j) = m_{ij}$  is the transition probability. Considering the transition probability from one node to another in  $t$  steps as  $M.M^{t-1}$ , the transition probability is inflated, i.e., higher transition probabilities are increased and lower transition probabilities are decreased. This is done by taking  $m_{ij}$  to the power  $r > 1$ , where  $r$  is an inflation parameter and defined using equation 1

$$\Upsilon(M, r) = \left\{ \frac{(m_{ij}^r)}{\sum_{a=1}^n (m_{ia}^r)} \right\}_{i,j=1}^n \quad (1)$$

The markov clustering method performs matrix expansion and inflation iteratively, i.e., it takes successive powers of  $M$  and then performs the inflation process. The iteration terminates when  $\|M_t - M_{t-1}\| \leq \epsilon$ , where  $\epsilon \geq 0$  is a threshold. In our study, we have used  $\epsilon = 0.001$  and the analysis is performed for different values of  $r$ . Table 8 and 9 gives the number of clusters and their size obtained at different values of  $r$  for Twitter and Facebook

Table 8: Twitter spam campaign analysis

Inflation parameter ( $r$ )	Total no. of clusters	Size (no.)
7	1	159(1)
10	25	12(12), 113(12), 34(1)
20	35	42(12), 2(2), 12(12), 95(1), 8(8)
35	39	12(24), 8(12), 117(1), 2(2)

spam profiles, respectively. We observe that for larger values of  $r$ , the number of clusters is increased, and this increase in the number of clusters is either due to the presence of outlier nodes or majority of overlapping nodes among clusters. Further detail about our spam campaigns analysis is presented in the following sub-section.

### 6.1. Cluster Analysis

After the application of the Markov clustering algorithm, we analyzed the spam profiles according to their respective clusters. We found that profiles in each cluster are connected to each other. Based on our observations of profiles in each cluster, we found 3 spam campaigns from our Twitter dataset, which consists of 5, 6 and 8 profiles. The campaigns with 5 and 6 profiles are part of the 8-node clusters. The 5-profile campaign utilizes the Twitter link sharing facility for carrying out advertisements of various products of single organization. However, the 6-profile campaign consists of profiles generating spam through automated activity because the tweets pattern generated by each profile is same. Moreover, the 8-profiles campaign is part of 12-node cluster which also exhibits similar behavior. In case of Facebook, we found 4 different spam campaigns consisting of 3, 5, 6 and 8 profiles. The 3-profile campaign is clustered separately and is an advertisement campaign. The 5-profile campaign consists of compromised accounts or accounts infected by spam-generating applications. On analysis we found that all the accounts are infected by the same application, generating spam posts without user permissions. The 6- and 8-profile campaigns consists of accounts under the control of a single spammer. These profiles spread spam by tagging community pages on spam posts.

## 7. Conclusion and Future Work

In this paper, we have proposed a set of 14 generic statistical features identified from Facebook and Twitter datasets to identify spam profiles on different types of social networks. We have also analyzed the discriminative



Table 9: Facebook spam campaign analysis

Inflation parameter ( $r$ )	Total no. of clusters	Size (no.)
2.0	4	6(1), 40(1), 95(1), 24(1)
2.2	5	6(1), 40(1), 94(1), 24(1), 1(1)
2.4	5	6(1), 38(1), 90(1), 24(1), 7(1)
2.6	7	7(1), 39(1), 90(1), 23(1), 1(1), 2(1), 3(1)

properties of the identified features using different categories of classification algorithms – naïve Bayes, Jrip, and J48. To this end, we have performed two different set of experiments. In the first experiment, the role of the whole feature set and their accuracy in terms of FPR and DR is judged over individual and combined datasets, whereas in the second experiment, we removed each feature successively and analyzed the results to judge the contribution of individual features towards spam profile detection. Our experimental result shows that the best classification result for the dataset containing both facebook and Twitter profiles can be achieved by using J48 algorithm. In this paper, we have also presented our study towards modeling social network using a weighted-graph and applying graph-based clustering method for spam campaign analysis. For this, we have considered the seven most discriminative features and identified different types of spam campaigns on Facebook and Twitter networks using Markov clustering (MCL) algorithm. At present, we are working towards identifying more features to increase spam profile and thereby spam campaign detection accuracy. We are also working towards the development of a crawler to enhance our dataset and to identify some other types of spam campaigns on Facebook and Twitter networks.

## Acknowledgment

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

## References

- [1] F. Ahmed, M. Abulaish, An mcl-based approach for spam profile detection in online social networks, in: Proceedings of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom'12), IEEE, 2012.

- [2] F. Benevenuto, T. Rodrigues, M. Cha, V. Almeida, Characterizing user behavior in online social networks, in: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, ACM, 2009, pp. 49–62.
- [3] L. Bilge, T. Strufe, D. Balzarotti, E. Kirda, 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, ACM, 2009, pp. 551–560.
- [4] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th International Conference on World Wide Web, ACM, 2011, pp. 675–684.
- [5] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, B. Zhao, Detecting and characterizing social spam campaigns, in: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, ACM, 2010, pp. 35–47.
- [6] C. Grier, K. Thomas, V. Paxson, M. Zhang, @ spam: the underground on 140 characters or less, in: Proceedings of the 17th ACM Conference on Computer and Communications Security, ACM, 2010, pp. 27–37.
- [7] J. Jiang, C. Wilson, X. Wang, P. Huang, W. Sha, Y. Dai, B. Zhao, Understanding latent interactions in online social networks, in: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, ACM, 2010, pp. 369–382.
- [8] X. Jin, C. Lin, J. Luo, J. Han, A data mining-based spam detection system for social media networks, Proceedings of the VLDB Endowment 4 (12).
- [9] E. Kartaltepe, J. Morales, S. Xu, R. Sandhu, Social network-based botnet command-and-control: emerging threats and countermeasures, in: Proceedings of the 8th International Conference on Applied Cryptography and Network Security (ACNS), 2010, pp. 511–528.
- [10] H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media?, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 591–600.

- [11] K. Lee, J. Caverlee, , S. Webb, Uncovering social spammers: Social honeypots + machine learning, in: Proceeding of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2010, pp. 435–442.
- [12] K. Lee, J. Caverlee, Z. Cheng, D. Sui, Content-driven detection of campaigns in social media, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM), ACM, 2011, pp. 551–556.
- [13] K. Lee, B. Eoff, J. Caverlee, Seven months with the devils: A long-term study of content polluters on twitter, in: Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM), 2011.
- [14] M. McCord, M. Chuah, Spam detection on twitter using traditional classifiers, in: Proceedings of the 8th International Conference on Autonomous and Trusted Computing, Springer, 2011, pp. 175–186.
- [15] S. Nagaraja, A. Houmansadr, P. Piyawongwisal, V. Singh, P. Agarwal, N. Borisov, Stegobot: a covert social network botnet, in: Proceedings of the 13th International Conference on Information Hiding, Springer, 2011, pp. 299–313.
- [16] A. Sala, L. Cao, C. Wilson, R. Zablit, H. Zheng, B. Zhao, Measurement-calibrated graph models for social network experiments, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 861–870.
- [17] G. Stringhini, C. Kruegel, G. Vigna, Detecting spammers on social networks, in: Proceedings of the 26th Annual Computer Security Applications Conference, ACM, 2010, pp. 1–9.
- [18] Symantec, Symantec Intelligence Report, August 2011.
- [19] K. Thomas, C. Grier, J. Ma, V. Paxson, D. Song, Design and evaluation of a real-time url spam filtering service, in: IEEE Symposium on Security and Privacy, 2011.
- [20] K. Thomas, C. Grier, V. Paxson, D. Song, Suspended accounts in retrospect: An analysis of twitter spam, in: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement, 2011, pp. 243–258.

- [21] K. Thomas, D. Nicol, The koobface botnet and the rise of social malware, in: Proceedings of the 5th International Conference on Malicious and Unwanted Software (MALWARE), IEEE, 2010, pp. 63–70.
- [22] S. van Dongen, Graph clustering via a discrete uncoupling process, Siam Journal on Matrix Analysis and Applications 30-1 (2008) 121–141.
- [23] A. Wang, Don't follow me: Spam detection in twitter, in: Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT), IEEE, 2010, pp. 1–10.
- [24] C. Yang, R. Harkreader, G. Gu, Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers, in: Proceedings of the 14th International Symposium on Recent Advances in Intrusion Detection (RAID'11), 2011.
- [25] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Zhao, Y. Dai, Uncovering social network sybils in the wild, 2011.