Spammer detection of social networking sites using 4 novel techniques

Sachin H. Bankar
ME-II Computer Engineering
Vidya Pratishthan's College of Engineering
Baramati, India
isachinbankar@gmail.com

Prof. S. A. Shinde
Department of Computer Engg.
Vidya Pratishthans College of Engineering
Baramati, India
chavanmanik@gmail.com

***Abstract*—Social Networking sites number of users are increasing day by day. Facebook is haing 2 billion users from all over the world. Twitter also having large number of users and they pay the dues if account gets hacked. Many researchers are working on to provide the securities to social networking sites. Providing security to social networking site has become a major task for researcher. Social networking sites have the personal data of users where they interact with each other. But spammers use these sites as platform to spread spam, malware and viruses. There are few techniques which are used to detect the spammers such as machine learning and honey-pot. Spammers are totally aware of these techniques, they know how to penetrate those security techniques and spread the spams. We are introducing novel technique which uses the graph generation, timed based, API based and neighbour based techniques to detect the spammers on social networking sites, which reduces the false positive rate and increases the spam detection rate.**

***Keywords*—SN's(Social networking Sites), spam, Twitter, Facebook, Machine Learning.**

I. Introduction

In the age of social media, according to a Wall Street Journal report, the microblogging service of Twitter spews out over 200 million tweets every day, having more than 190 million accounts[2].As we know that it is now part of a daily routine of peoples life. People share their photos, videos and experiences and also they are able to track the people and hottest trends.[1]

Another SN's which gaining more and more popularity is Facebook which helps us to share videos, photos and updates about our personal life. On Facebook 1.19 billion users are active monthly as of 30, September 2013. Everyday data increases by 1.5 petabyte.

SN's has became a highly attraction for spammers to spread the malwares and achieve their malicious goal by sending spam,[7] hosting botnet command and control (C and C)channels [5]

In new approach adding 10 new features to an existing SN's Spammer Detection Techniques to increase the detection of spammers and robustness of detection scheme. Spammers and researchers are competing with each other from the beginning of the Social networking site existence.

As number of spammers increases most of the researchers devoted their life for spammer detection. SN's contain large number of users, Status updates, URLs, so as time passes SN's spammer learn to evade the spammer detection techniques, such as Profile Based, Content based etc. In new approach evaluate and analyse the detection technique and implement 10 new features to detect them. New features are based on Graph, Neighbour, Time and Automation. And it also increases the robustness of the detection scheme.

II. Literature Survey

A. Creating Hoenypot[2]

In honeypot based approach, honeypot is deployed over the social networking sites and attract the spammers. Then classify the features for the detection of spammers. These features are used to detect the spammers which are new and also existing spam accounts.
Advantages:

1) Honeypots can give you the exact information you need in a format which is quick and easy.
2) Reduction of False Positives and False Negatives.
3) Simplicity: Easy to enhance the security in any organisation because of simplicity.

Disadvantages:

1) Single Data Point: Honeypots are of no use if no one attacts it.
2) Risk: It may risk your environment.
3) Continuous administration is necessary for Honeypots to give the required output.

B. Machine Learning:[3]

Collects the data from the social networking sites such as users, tweets etc. Then separate them as spammers and non-spammers account. Find out the features of the spam accounts which can be used to detect the spammers on social networking site, and implement them to detect the spam accounts.
Advantages:

1) False positive or negative rate is low.
2) Simple to implement

Disadvantages:

1) Required large number of data
2) It can be evade by sing new techniques

C. Content based and Profile based:[4]

Content based features such as their tweets and number of URLs duplication and repetition are used to detect the spammers. Profile based features are used to detect the spammers using the ”Followers” and ”sollowings” ratio.
Advantages:

1) the proposed reputation feature has the best performance of detecting abnormal behaviors
2) Large number of spams are detected

Disadvantages:

1) False positive rate is high.
2) Classification algorithm matrix is confusing.

D. Analysis of Spammers Behaviour:[5]

Proposed system analyse to which extent spam has entered social networks. More precisely, proposed system analyse how spammers who target social networking sites operate. To collect the data about spamming activity, ”honey-profiles” collect the data and tweets and log the data and tweets. Then analyse the collected data and identify anomalous behaviour of users who contacted honey-profiles. Based on the analysis of this behaviour, new techniques are used to detect spammers in social networks, and then aggregate their messages in large spam campaigns. Proposed systems results show that it is possible to automatically identify the accounts used by spammers, and our analysis was used for take-down efforts in a real-world social network.

III. Implementation Details

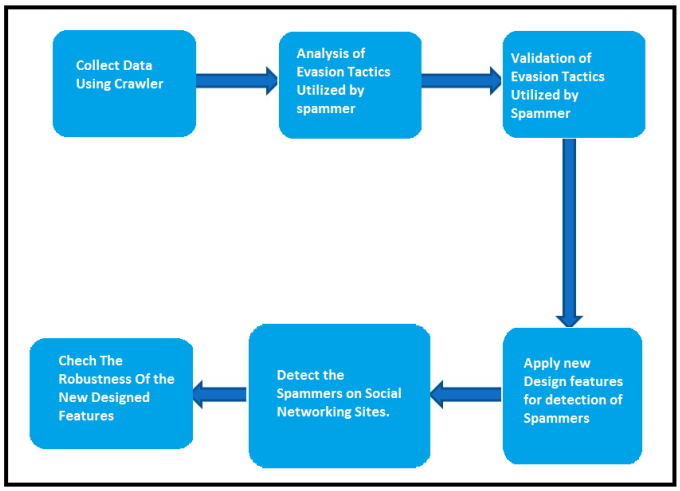

Fig. 1. Architecture of the Proposed System

A. Data Collection

To crawl social networking sites profiles, use a crawler that taps into social networking sites Streaming API. In order to decrease the effect of possible sampling bias, our crawler recursively collects accounts data in multiple rounds, with the consideration of guaranteeing sampling randomness and maintaining social relationships, rather than simply using the Breath First Search (BFS) sampling technique. More specifically, in each round, our crawler first collects 20 seed social networking sites accounts from the public timeline, which are randomly selected by sites [7]. Then, the crawler will collect all of those seed account's followers and followings. This crawling process will be repeated in the next round. Also, for each account, our crawler collects its 40 most recent Tweets and the URLs in the tweets. Due to the large amount of redirection URLs used in sites, we also follow the URL redirection chain to obtain the final destination URL.

B. Analysis of Evasion Tactics

1) **Profile Based Evasion**
   - **Gaining More Followers** To Evade the profile based technique spammers gain the more number of followers or we can say they buy the followers. That's the reason it is difficult for detection scheme to detect them.
   - **Posting More Tweets** Posting more tweets increases the active time of an spammer on the twitter and also increased the reputation score of an spammer. So posting more and more tweets they can evade the detection scheme.
2) **Content based Evasion**
   - **Mixing Normal Tweets** Spammers while posting the malicious Tweets also post the normal tweets to evade the detection scheme. Instead of continuously posting URLs and Advertisement it also post the Tweets which are not relate to the malicious behaviour.
   - **Posting Heterogeneous Tweets** Previous scheme used to detect the spammers using the number of similar Tweets, but spammers now use the automation system to evade the detection scheme which is having similar meaning but with different wordings.

C. Applying New Features

1) **Graph-Based Features:**
   a) **Local Clustering Coefficient**
   Since legitimate users usually follow accounts whose owners are their friends, colleagues or family members, these accounts are likely to have a relationship with each other. However, since spammers usually blindly follow other accounts, these accounts usually do not know each other and have a looser relationship among them. Thus, compared with the legitimate accounts, Twitter spammers will have smaller local clustering coefficients.

$$LC(v) = \frac{2|e^v|}{K_v.(K_v - 1)} \quad (1)$$

Here,
v-Vertex
e^v-the total number of edges built by all vs
K_v-the sum of total indegree and outdegree of the vertex v

b) **Betweenness Centrality** A Twitter spammer will typically use a shotgun approach to finding victims, which means it will randomly follow many unrelated accounts. As a result, when the Twitter spammer follows these unrelated accounts, the spammer creates a new shortest path between those accounts through the spam account. Thus, the betweenness centrality of the spammer will be high.

$$BC(v) = \frac{1}{(n-1)(n-2)} \cdot \sum_{s \neq v \neq t \epsilon V} \frac{\delta_{st}(v)}{\delta_{st}} \tag{2}$$

6 Here,
δ_{st}-The number of Shortest Path from s to t
$\delta_{st}(v)$-The number of Shortest Path from s to t that passes through a vertex v
n-The total number of Vertex

c) **Bidirectional Links Ratio** If two accounts follow each other, we consider there is a bidirectional link between them. The number of bidirectional links of an account reflects the reciprocity between an account and its followings. Since Twitter spammers usually follow a large number of legitimate accounts and cannot force those legitimate accounts to follow back, the number of bidirectional links that a spammer has is low. On the other hand, a legitimate user is likely to follow his/her friends, family members, or coworkers who will follow this user back. Thus, this indication can be used to distinguish spammers.

$$R_{bilink} = \frac{N_{bilink}}{N_{fing}} \tag{3}$$

Here,
N_{bilink}-Total Number of bidirection Link
N_{fing}-Total number of Following

2) **Neighbour-Based Features:**
As we know that spammer can change their behaviour but they don't have any control on others accounts behaviour. Using this feature we can distinguish the spammers from legitimate accounts.

- **Average Neighbour's Followers:** Legitimate typically follow the accounts which are quiet quality accounts. Using these features calculate the neighbours average followers and then distinguish the spammers from legitimate users.

$$A_{nfer}(v) = \frac{1}{|N_{fing}(v)|} \cdot \sum_{u \epsilon N_{fing}(v)} N_{fer}(u) \tag{4}$$

Here,
N_{fer}-Number of Followers
N_{fing}-Number of Followings

- **Average Neighbour's Tweets:** As the above also calculating the average of tweets of neighbours can help us to find the suspicious accounts. Same formula i.e. above mentioned can be implemented to calculate the Average Neighbours Tweets.
- **Followings to median neighbours followers:** To extract this feature, we first calculate the median number of an accounts all following accounts follower numbers.

$$R_{fing_mnfer} = \frac{N_{fing}}{M_{nfer}} \tag{5}$$

3) **Automation-Based Features:**
Due to the high cost of manually managing a large number of spam accounts, many spammers choose to create a custom program using Twitter API to post spam tweets. Thus, we also design three automation-based features to detect spammers: API Ratio, API URL Ratio and API Tweet Similarity.

- **API Ratio:** The ratio of the number of tweets with the tweet source as API to the total number of tweets. As existing work shows, many bots use API to post tweets, so a higher API ratio implies this account is more suspicious.
- **API URL Ratio:** The ratio of the number of tweets containing a URL posted by API to the total number of tweets posted by API. It is more convenient for spammers to post spam tweets using API, especially when spammers need to manage a large amount of accounts, as discussed in Section IV. Thus, a higher API URL ratio of an account implies that this accounts tweets sent from API are more likely to contain URLs, making this account more suspicious.
- **API Tweet Similarity:** Spammers can use tricks to evade the detection feature of tweet similarity and still choose to use API to automatically post malicious tweets. Thus, we also design API tweet similarity, which only compute the similarity of those tweets posted by API. Thus, a higher API tweet similarity of an account implies that this account is more suspicious.

4) **Timing-Based Features:**
Timing based feature nothing but an tweeting rate and at the sane time following rate of an account.

- **Following Rate:** Reflects the speed at which an account follows other accounts. Since spammers usually follow many users in a short period of time, a high following rate of an account indicates that the account is likely a spam account.

D. Results

Table shows that for each classifier, with the addition of our newly designed features, the detection rate (DR) increases over 10%, while maintaining an even lower false positive rate (FPR). This observation validates that the improvement of the

detection performance is indeed due to our newly designed features.

	Without Our Features			With Our Features		
Classifier	FPR	DR	F-1 Measure	FPR	DR	F-1 Measure
Random Forest	0.013	0.737	0.791	0.004	0.848	0.9
Decision Tree	0.014	0.697	0.760	0.008	0.840	0.876
BayesNet	0.068	0.762	0.629	0.01	0.838	0.833
Decorate	0.012	0.697	0.768	0.012	0.854	0.884

Fig. 2. Results of new detection technique and comparison with previous techniques

E. Future works

Collecting a large number of ideal dataset is practically impossible. Our crawled dataset may still have a sampling bias, and Twitter and facebook dataset without any bias is hardly possible.

In addition, it is well acknowledged in the community that it is challenging (or impossible) to achieve a comprehensive ground truth for social networking sites spammers. Also, in order to guarantee that our collected spammers are real spammers, we use a more strict strategy than what used in most of other related work to collect our spammers. Thus, the number of our identified spammers is only a lower bound, and the percentage of identified spammers in our dataset may be smaller than that reported in other studies. However, even for a subset of spammers, we can see that they are evolving to evade detection. And our evaluation validates the effectiveness of our newly designed features to detect these evasive spammers. We also acknowledge that some identified spam accounts may be compromised accounts. However, since these accounts still behave fairly maliciously in their recent histories, it is meaningful to detect them.

We clearly admit that those 20K accounts used as our benign dataset may still contain some spam accounts. However, it is very difficult to obtain a perfect ground truth from such a big dataset. Thus, we only collect those accounts without posting malicious URLs to build the benign dataset. Also, we believe that our major conclusion could still be held, although there could be some noisy items in the training dataset.

While graph-based features such as local clustering co-efficient and betweenness centrality are relatively difficult to evade, these features are also expensive to extract. Also, precisely calculating the values of such graph metrics on large graphs (e.g.,the whole Twitter graph) is very challenging and a hot research issue,which is out of scope of this work. However, we could still estimate the values of these two features by using a neighbour sampling technique that allows us to compute these metrics piece-by-piece. Also, since we can not extract the exact time when an account follows another, we use an approximation to calculate the feature of following rate. Even though this feature may be not perfectly accurate, an approximate value of this feature can still reflect how radically an account increases its following number.

For future work, we plan to design more robust features, evaluate our machine learning detection scheme on larger datasets by using more crawling strategies, and work directly with social networking sites. We also plan to broaden our targeted type of spammers, so that we can perform a deeper analysis on the evasion tactics by different types of spammers. We also plan to make more quantitative models for the analysis of the robustness of the detection features by deeper analysing the envision tactics. In addition, further studies on analysing the correlation among different features and designing better machine learning classifiers by selecting more effective features are also in our future plan.

F. Conclusion

To detect the spammers on social networking sites, we proposed a novel 4 state art detection scheme, which are having features like graph based, neighbour based, automation based and time based . Through the analysis of those evasion tactics and the examination of four state-of-the-art solutions, we design several new features. In addition, in terms of spammer's dual objectives staying alive and achieving malicious goals, we also formalize the robustness of detection features for the first time in the literature. Finally, according to our evaluation, while keeping an even lower false positive rate, the detection rate by using our new feature set is also much higher than all existing detectors under four different prevalent machine learning classifiers.

REFERENCES

[1] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, Measuring User Influence in Twitter: The Million Follower Fallacy, in Proc. Int. AAAI Conf. Weblogs and Social Media (ICWSM), Washington, DC, USA, 2010.

[2] K. Lee, J. Caverlee, and S.Webb, Uncovering social spammers: Social honeypots machine learning, in Proc. ACM SIGIR Conf. (SIGIR), Geneva, Switzerland, 2010.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, Detecting spammers on Twitter, in Proc. Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS), Redmond, WA, USA, 2010.

[4] A. Wang, Dont follow me: Spam detecting in Twitter, in Proc. Int. Conf. Security and Cryptography (SECRYPT), Athens, Greece, 2010.

[5] G. Stringhini, S. Barbara, C. Kruegel, and G. Vigna, Detecting Spammers on social networks, in Proc. Annual Computer Security Applications Conf. (ACSAC10), Orlando, FL, USA, 2010.

[6] The Twitter Rules, 2011 [Online]. Available: http://help.twitter.com/entries/18311-the-twitter-rules

[7] Auto Twitter, 2011 [Online]. Available: http://www.autotweeter.in/

[8] The 2000 Following Limit Policy On Twitter, 2009 [Online].Available: http://twittnotes.com/2009/03/2000-following-limit-on-twitter.html

[9] Twitter API in Wikipedia, 2011 [Online]. Available: http://apiwiki.twitter.com/