Ecosystem of Spamming on Twitter: Analysis of Spam Reporters and Spam Reportees

Pooja Sinha, Oshin Maini, Gunjan Malik and Rishabh Kaushal
Department of Information Technology
Indira Gandhi Delhi Technical University for Women, Delhi, India

***Abstract*—Lately, there has been a growing trend in the Internet space particularly among the Online Social Media (OSMs) platforms like Twitter, Facebook etc which are becoming huge repositories of information. This information, by design, is posted by users of these websites and consequently, this information is vast, un-organized, unreliable and dynamic. It is commonly observed that along with genuine users, a lot of activity is seen from spammers or users with the intent of spreading malicious or irrelevant content. In our work, we focus on spamming activity on Twitter. Spamming activity in Twitter is can typically be reported by its users, who we refer as *reporters* and those who indulge in spamming activities are referred as *reportees*. We collected data of suspected spammers, i.e. reportees as well as of the users who reported them, i.e. reporters. Thereafter, we classified them into various categories and tried to study the ecosystem of these reportees and reporters. We used three data mining techniques i.e., decision tree , K-nearest neighbors and random forest classifier for the classification tasks. Finally, we have compared these three algorithms on the basis of their accuracy.**

Keywords*—*Twitter, Spamming, Data Mining, Decision Tree Classification, K-Nearest Neighbors, Random Forest Classifier.

I. Introduction

Online Social Media (OSM) platforms are becoming increasingly popular among Internet users. By design, these websites grow by giving their users ease of posting content related to their personal life or any other event or phenomena in social, economic, business, films or any other. Consequently, a vast amount of information is hosted on these websites or platforms. Our work, through this paper, is an extensive research on spamming on a popular social networking website i.e. Twitter which is a micro blogging social networking platform allowing its users to post small messages (*tweets* of 140 characters or less). By default, the content posted by users is *public* so that it has wide visibility. A user can *follow* any user; thereby obtaining all the messages posted by them. The other user may not necessarily follow back that user. This way the spread of information happens at a very fast pace. A user may *retweet* (re-post) any message that she has received from the users that she is following. Also, a user may direct a message to another user either privately or publicly using (*mentions*). Twitter also allows users to organize their content by categorizing them under a common identifier (*hashtag*). The social platform provides a lot of freedom to its users to disseminate information over the social network. As is observed that this freedom is misused by few users who post inappropriate and irrelevant content as well, referred as *spamming*. There is a need to put a stop to such malicious activities by these malicious users, who are referred as *spammers*. To address the issue, Twitter has provided an option for genuine or normal users, wherein they can report a spammer manually through the option provided on the user's profile or even a tweet can be reported. It has also come to our notice that users tweet using *@spam* mention followed by the suspected spammer's userID through *@username* mention. For instance, as we can see in Figure 1, a user 'kurukuru' (reporter) is reporting another user '1000favs' (reportee) as suspected spammer.

Fig. 1. An indicative snapshot depicting reporting of spammers (Reportees) on Twitter by users (Reporters) using *@spam* mention in their tweet

Twitter itself checks the reported spammers and if correctly reported, these user accounts are suspended. Twitter defines spamming as the peculiar kind of user behavior or tweets that do not follow Twitter guidelines. Most common spamming behaviour that we observed on Twitter are in one of the following forms (1) Automated accounts posting inappropriate content. (2) Automated accounts for aggressive following or aggressively tweeting posts on trending topics. (3) Many passive automated accounts with similar Twitter handles. (4) Automated accounts posting publicity related tweets.

There have been many research studies conducted to automate spammer detection but most of them focus on classifying a user as either a spammer or a genuine user. Our study aims at classifying the suspected spammers as well as the users that reported these spammers into different categories on the basis of their behavior on the network. All of the spam detection algorithms presently focus only on the attributes of the suspected spammer when trying to identify whether the

978-1-5090-2029-4/16/$31.00 @2016 IEEE

user is a spammer or not. In our view, an algorithm that focuses on the attributes of the person who is reporting the spammer as well would have a greater accuracy in correctly identifying spammers and would prevent genuine users from getting suspended.

This paper is organized as follows. We briefly mention related research in next Section which is followed by description of data and features. Thereafter, we explain our implementation, present results and finally conclude the work.

II. Related Work

Wang et. al. [1] proposed a directed social graph model wherein they describe a follower and friend model. Based on the spam policy of Twitter, content-based and graph-based features were proposed in the paper. Traditional classification algorithms are applied on the dataset and graph generated data to detect suspicious behaviors of spam accounts. Their results show that the Bayesian classifier has a better overall performance and achieve an 89% precision. McCord et. al. [4] studied several attributes useful to differentiate spammers and non-spammers. They suggest several user-based and content-based attributes to make this distinction. They evaluate the usefulness of these attributes in spammer detection using traditional classifiers like Random Forest, Nave Bayesian, Support Vector Machine, K-nearest neighbor schemes. Among the four classifiers they evaluated, the Random Forest classifier produces the best results. The selection of classifiers for our research is inspired from the approach used in Gupta et. al. [3]. They have used Nave Bayes classifier, k-NN classifier and Decision Trees to classify the data into spammers and non-spammers. They first compared the result of Nave Bayes and k-NN classifier.

Although, broadly speaking, we too have used the approach of using well known classifiers, however, we apply them to classify reporters and reportees in the ecosystem of spamming activity details of which shall be discussed in next sections.

III. Data Collection

For data collection in Twitter, we used python as our implementation language and *tweepy python* wrapper to make calls to fetch data from the Twitter API. Applications were created to generate tokens that would help us gain access to the API. A python script was then written using this wrapper to collect users' data and compute the features that we shall describe in next section. Since we have chosen one specific kind of tweeting pattern for reporting spammers as the basis of data collection, we collected tweets that contained *@spam* as a substring. We collected a dataset of over 8700 tweets. The user who has published the tweet is the reporter while the user mentioned in the tweet is the reportee. Using automated means, we segregated the tweet data into reporters and reportees. A dataset of 384 reportees and 96 reporters was extracted from these tweets. On the basis of these metrics, both reporters and reportees were classified into the identified categories. To compare the behavior of reportees with genuine users, we added 135 genuine users in our dataset. We also added around 100 promoters to be able to understand the promotion related behavioral activities.

IV. Categorization and Feature Set Description

We examined the accounts of several reportees and identified the following classes of reportees:
Active Spammers: These reportees are very active on Twitter and their most recent tweets have been posted less than a week ago. These users are involved in posting malicious and irrelevant content on Twitter including inappropriate pictures and malicious URLs. These accounts have a very small follower-followee ratio and a high number of hashtags, mentions, spam words, repeated tweets and links.
Promoters: The users in this class are also very active on Twitter but instead of posting inappropriate content, they post content that publicizes themselves. They may be a company, an organization, celebrities or users that are very popular on Twitter. They have a stable follower-followee ratio. They might have hashtags and mentions but they are less in number as compared to suspected spammers.
Genuine: These are genuine users on the network that have wrongly been reported by some other user. These users tweet relevant content on Twitter and have no intentions of spreading malicious content. They have a high follower-followee ratio and a moderate number of hashtags, links and mentions and practically no spam words or repeated tweets. *Passive Spammers*: These are users that might have all or some of the characteristics of spammers. But the attribute that makes them passive is their last tweet date. If its a long time ago, they are passive spammers.

After examining the accounts of the reporters that we collected, we were able to identify the following classes of reporters:
Automated Account: While looking at the accounts of various reporters, we observed a peculiar behavior in the accounts of some of these reporters. These users posted only @spam @username tweets. They were otherwise inactive on the network i.e. they had almost no friends and followers. However, the number of tweets that they had posted till date was very large.
Genuine Users: These are genuine users with the intent of using Twitter to post genuine content. They are neither very popular like celebrities nor are using Twitter for promotional or publicity purposes.
Popular/Promotional/Celebrities: These accounts generally have features like large following and large number of tweets. Celebrities, news channels, organizations, popular users on Twitter or other online social media, also known as internet celebrities fall under this category.

Figures 2 and 3 depict hashtags and mentions per tweet for the reporters that we collected. Number of hashtags per tweet is found to be less for genuine users than promoters.

In order to automatically classify these identified reporters and reportees into these classes, we need metrics or attributes or features on the basis of which such classification could be done. For identifying the features, we have referred some of the earlier works ([1], [2], [4]). Feature set description for reportees are given in Table I and feature set description for reporters is given in Table II.

Table III gives the mean values for the metrics that were extracted for Reporters.

After we identified these metrics, we manually inspected

TABLE I. FEATURE SET DESCRIPTION FOR REPORTEES

Feature	Description	Reason for Use
Reputation	It is the ratio of the number of followers and the sum of followers and followees of a user.	For any spammer, this value is always very close to zero. It has been observed in a survey of Twitter that the reason behind this is that the number of followers of a spammer is very small as compared to the number of people that they follow.
Hashtags	We have calculated the number of hashtags in the most recent 50 tweets of the user	We have already mentioned that hashtags are used to organize the content by placing them under a common header/identifier. Since genuine users aren't in the business of grabbing too much attention, they use lesser number of hashtags than spammers and promoters or celebrities.
Mentions	We have calculated the number of mentions in the most recent 50 tweets of the user.	Spammers tend to use a lot of mentions so as to catch the attention of a high number of users while genuine users use it to tag their friends and promoters don't use mentions as often. Hence, all these classes of twitter users use mentions differently
Spam Words	We have collected a dataset of 291 spam words by manual inspection. We have calculated the number of spam words in the most recent 50 tweets of the user.	Spammers use a set of words in spam tweets. These words are inappropriate and irrelevant and are called spam words. The rest of the reportees don't really use any of these words.
Links	We have calculated the number of links in the most recent 50 tweets of the user.	Since the size limit of a tweet is very small i.e., 140 characters, spammers are in the habit of directing users to other spam pages i.e. malicious pages or pages with unrelated content via hyperlinks. The same technique is used by promoters who use these links to direct users to their legitimate websites while genuine users include URLs in their tweets less frequently.
Account Age	It is the number of days since the date on which the Twitter account has been created	It has been observed that recently created accounts turn out to be those of spammers as their accounts keep getting detected and suspended.
Number of Days since last tweet	It is the number of days since user's most recent tweet	Number of Days since last tweet helped us check the passivity of users. Passivity refers to the frequency of a users tweets. Passive spammers have their last tweet date around 2 - 3 years ago while genuine users, active spammers and promoters have a more recent last tweet date. This metric has basically been used to discriminate between passive and active users.
Number of tweets	This is the number of tweets that the user has posted till date	This metric is used to judge how active the user is on Twitter.
Minimum Interval Between Tweets	This metric is the minimum of the time differences between the consecutive tweets on the day on which the user has posted maximum number of tweets.	We observed that most of the spammers have very little time differences between consecutive tweets. These might even be less than a minute. Also, they usually post a lot of tweets within a few hours on a particular day rather than spamming every day. This might also happen in the case of promoters but promoters are generally popular on the network and have a high following.

TABLE II. FEATURE SET DESCRIPTION FOR REPORTERS

Feature	Description	Reason for Use
Number of followers	The number of followers that a Twitter user has	This metric helped us determine the popularity of the reporter on the network.
Number of followees	The number of Twitter accounts that a user follows	A celebrity has more followers than followees. Automated accounts have less than 10 users following them. There is a difference in the number of followees of each category of reporters.
Number of reports	The number of Twitter accounts that the user has reported as spam	Automated accounts have been observed to be tweeting only spam reports while other classes of reporters have fewer reports amongst their tweets as compared to the automated ones.
Tweet Count	Total number of tweets that a user has posted till date	It helps judge how active a user is on Twitter
Retweets	The number of tweets of a reporter that are retweets	For automated accounts, this number is almost zero while it is higher for genuine users and very high for promoters and celebrities
Retweeted	The number of tweets of a reporter that have been retweeted	For automated accounts, this number is almost zero while it is higher for genuine users and very high for promoters and celebrities.
Mentions	Percentage of mentions in the tweets.	This number is very high for automated accounts as all their tweets are reports and contain mentions of various spam accounts. Genuine users and promoters have fewer mentions amongst their tweets.
Hashtags	The number of hashtags that the user has in all his tweets	This along with other metrics helps check whether the reporter himself is not displaying spammer tendencies.

and classified all the collected users (reporters as well as reportees) into the categories that were identified in Table II and Table III, respectively above.

V. PROPOSED APPROACH

Implementation pipeline of our proposed approach is depicted in Figure 4 which shows the data collection, categorization of data and finally applying data mining algorithms on the computed features. For data collection, as mentioned earlier, a python script was implemented using the *tweepy* python wrapper and was also used to extract and compute metrics for the Reporters and the Reportees. We have mined data using Decision Trees, K-Nearest Neighbors and Random Forest Classifier. We have used R language to implement these data mining algorithms. A part of the dataset is fed as Training Data to the algorithm and the rest as Test data in the ratio of 2:1. The output of this algorithm was verified against manual classification to check the accuracy of the algorithms. The performance of these algorithms was then compared which is presented in next section.

TABLE III. MEAN VALUES OF FEATURE SET FOR REPORTEES

Category	Reputation	Hashtags	Mentions	Spam Words	Links	Account Age	# Days since last Tweet	# Tweets	Min. Tweet Freq
Genuine Users	0.31	0.44	0.77	0.04	0.60	793.78	92.02	8562.85	1476.82
Passive Spammers	0.06	6.65	0.36	2.40	0.74	793.72	1018.28	14514.60	107.68
Promoter/Celebrity	0.90	0.60	0.83	0.03	0.79	2001.44	12.58	51477.19	1.09
Active Spammers	0.04	8.45	0.31	4.02	0.85	735.80	128.56	1860.39	1.78

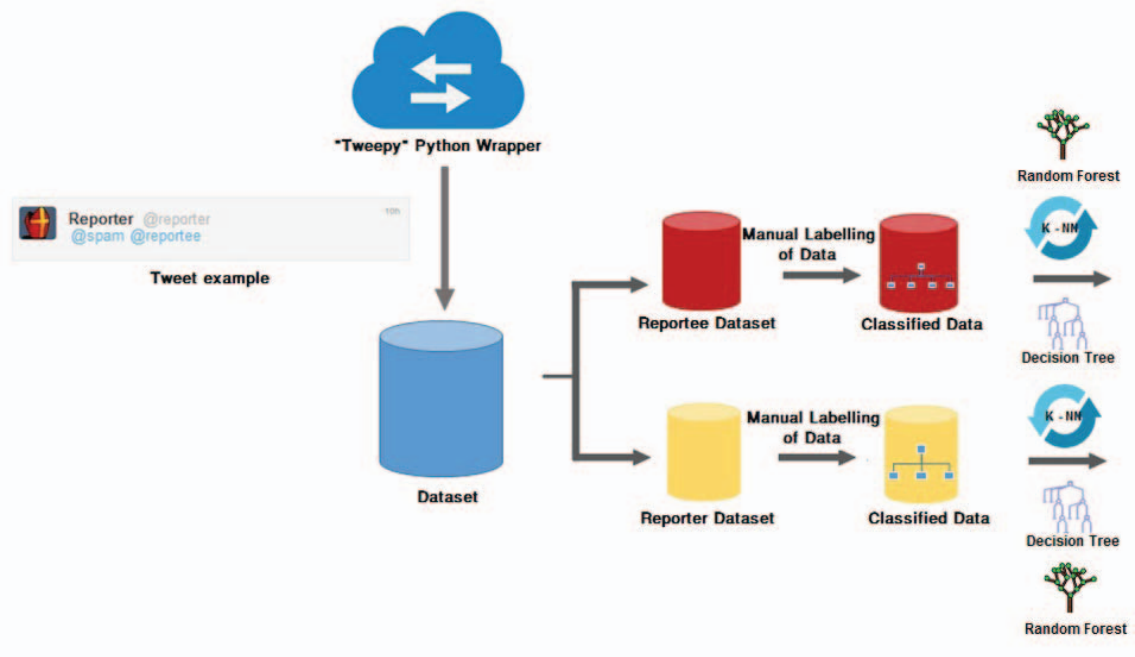

Fig. 4. Implementation Pipeline of Proposed Approach comprising of Data Collection, Labelling of Data and Application of Data Mining Algorithms

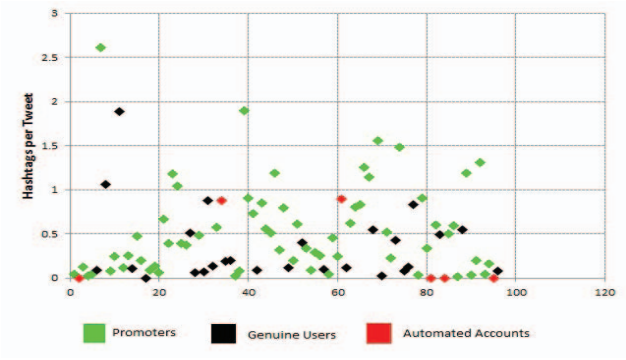

Fig. 2. Hashtags per Tweet of Reporters

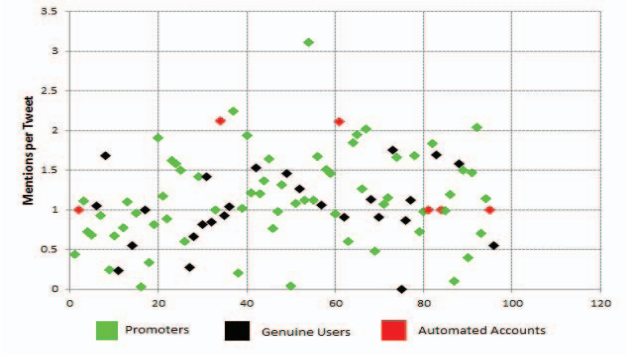

Fig. 3. Mentions per Tweet of Reporters

VI. RESULTS AND OBSERVATIONS

In the case of reportees, we chose K as 19 because we had a lot of users per class and hence each cluster had a lot of nodes close it, therefore, a larger K-value would help in reducing the noise. In the case of reporters, we chose K as 2 because a few of the classes of reportees had very less users and hence, to reduce noise, we had to choose a smaller value of K. In addition to KNN classifier, we have used decision trees and random forest classifier for classification of our data. We used a set of measures, to measure efficiency, commonly used in machine learning and information retrieval. We have considered the following measures: precision, recall and F-measure. A comparison of these measures for KNN classifier, decision trees and random forest classifier for reportees is given in Table IV while for reporters it is given in Table V. So from the values, it is quite clear that decision tree is the better classifier for reportees and reporters.

Figure 5 and Figure 6 show the manual and automated categorization by KNN classifier of reportees and reporters, respectively.

VII. NETWORK ANALYSIS

A reporter-reportee graph represented in Figure 7 was generated using Gephi in order to better understand the reporting phenomenon. As can be seen from the graph a large cluster of users is visible at the center. This represents automated accounts and the users that they report. Other clusters can be

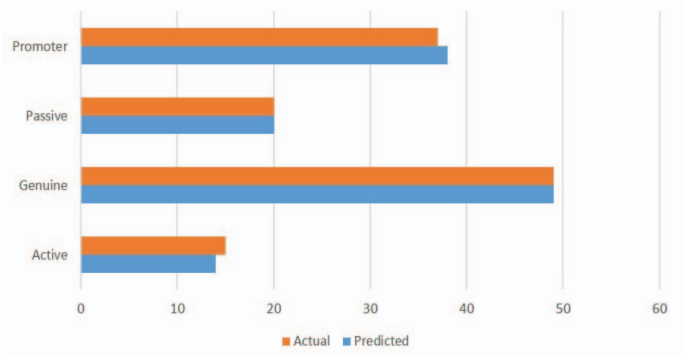

Fig. 5. KNN Classification for Reportees

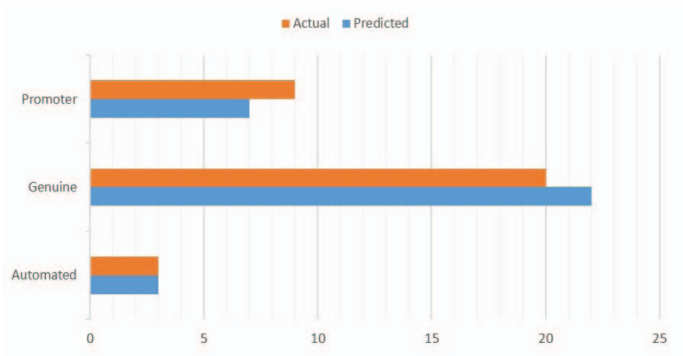

Fig. 6. KNN Classification for Reporters

TABLE IV. COMPARISON OF ACCURACY OF DECISION TREES, K-NN CLASSIFIER AND RANDOM FOREST CLASSIFIER FOR CLASSIFICATION OF REPORTEES

Category	Precision			Recall			F-Measure		
	K-NN K=19	Decision Tree C5.0	Random Forest	K-NN K=19	Decision Tree C5.0	Random Forest	K-NN K=19	Decision Tree C5.0	Random Forest
Genuine User	0.8	0.87	0.93	0.86	1	0.93	0.83	0.93	0.93
Passive Spammer	0.9	0.88	0.9	0.9	0.96	0.92	0.9	0.92	0.91
Promoter/Celebrity	0.9	1	0.95	0.9	1	0.9	0.9	1	0.92
Active Spammer	0.89	0.97	0.89	0.87	0.84	0.89	0.88	0.9	0.89

TABLE V. COMPARISON OF ACCURACY OF DECISION TREES, K-NN CLASSIFIERAND RANDOM FOREST CLASSIFIER FOR CLASSIFICATION OF REPORTERS

Category	Precision			Recall			F-Measure		
	K-NN K=2	Decision Tree C5.0	Random Forest	K-NN K=2	Decision Tree C5.0	Random Forest	K-NN K=2	Decision Tree C5.0	Random Forest
Automated Accounts	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Genuine Users	0.85	0.92	0.9	0.77	0.96	0.85	0.81	0.94	0.87
Promoters/Celebrities	0.45	0.75	0.7	0.57	0.60	0.78	0.50	0.67	0.74

found as well and depending on it's size we can gauge how actively that reporter reports other users.

VIII. CONCLUSION & FUTURE WORK

In this research, we have mainly focused on spamming as a malicious activity and finding the users(reporters and reportees) involved in it. As can be seen from our results, decision trees has correctly identified 92% of the time for reportees and 90% of the time for reporters, K-NN correctly classified 84% of the time for reporters and 89% of the times for reportees while random forest classifier correctly identified 90% of the time for reportees and 84% of the time for reporters. There are certain accounts that are automated which generate only spam reports. These accounts make for a lot of the spam reports that are generated.

An algorithm that takes into account this classification for reportees and their behavior with the people who report them, can be developed in the future for classifying spammers with greater accuracy. It can allocate points to each reportee based on a behavioral analysis of the reporter and the reportee. Attributes like who is following whom could be used to allocate these points. For example, if the reporter is following the reportee then positive could be given, else negative could be given. Based on such allotment of points, a sum total of points would be calculated for each reportee. If the final score is below a certain threshold then the reportee is a spammer,

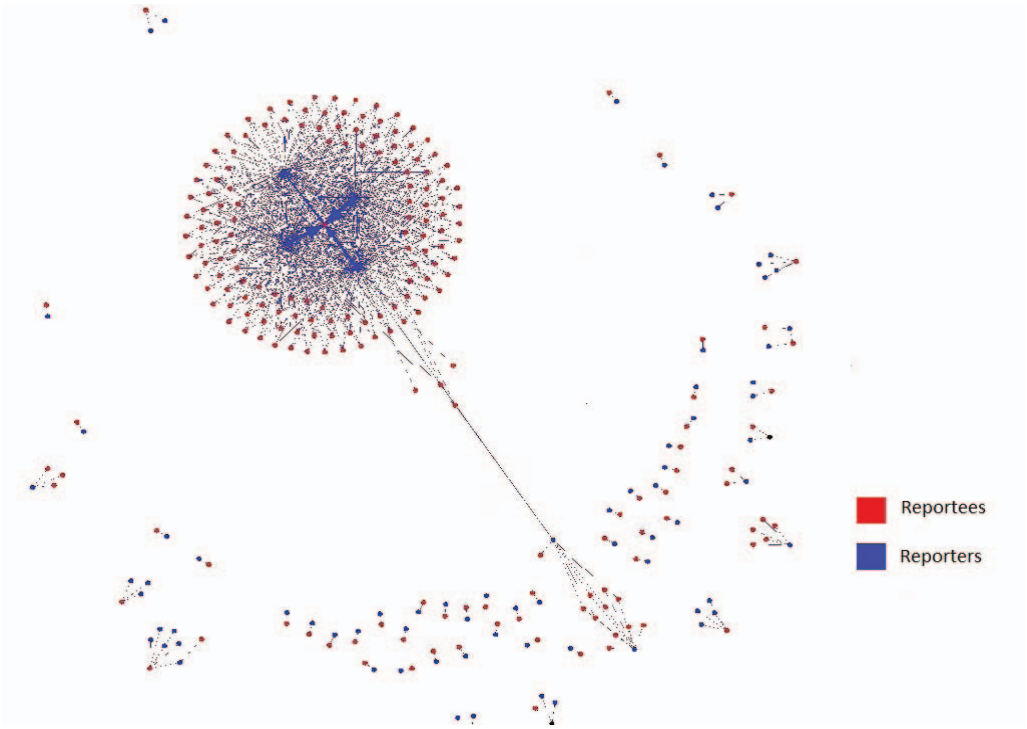

Fig. 7. A reporter-reportee graph depicting cluster of users at the center

else not.

REFERENCES

[1] A. Wang. *Dont follow me: Spam detection in Twitter*. in Proceedings of the International Conference on Security and Cryptography (SECRYPT 2010), 2010, Cited 184.

[2] M. McCord, M. Chuah. *Spam detection on Twitter using traditional classifiers*. Proceedings of the 8th International Conference on Autonomic and Trusted Computing, Springer, 2011, Cited 69.

[3] Gupta, Arushi, and Rishabh Kaushal. *Improving Spam Detection in Online Social Networks*.

[4] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. *Detecting Spammers on Twitter*.In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), July 2010, Cited 326

[5] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Zhao. *Detecting and characterizing social spam campaigns*, Proceedings of the Internet Measurement Conference (IMC), 2010.

[6] De Wang, DaneshIrani, and Calton Pu. *A Social-Spam Detection Framework*. , Proceedings of Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), 2011.

[7] J. Song, S. Lee, and J. Kim. *Spam filtering in Twitter using sender-receiver relationship*. In Proceedings of International Symposium on Recent Advances in Intrusion Detection (RAID), 2011, Cited 71

[8] Kurt Thomas, Chris Grier, Vern Paxson and Dawn Song. *Suspended Accounts in Retrospect: An Analysis of Twitter Spam*. Internet measurement conference (IMC), 2011.

[9] Dewan Md. Farid, Nouria Harbi and Mohammad Zahidur Rahman. *Combining Naive Bayes And Decision Tree For Adaptive Intrusion Detection*. International Journal of Network Security & Its Applications (IJNSA), Volume 2, Number 2, April 2010

[10] K. Yoshida, F. Adachi, T. Washio, H. Motoda, T. Homma, A. Nakashima, H. Fujikawa, and K. Yamazaki *Density-based spam detector*. Proceedings of the Tenth ACM SIGKDD International Conference, 2004.