Enhancing Twitter Spam Accounts Discovery Using Cross-Account Pattern Mining

Ioana-Alexandra Bara, Carol J. Fung, Thang Dinh
Department of Computer Science,
Virginia Commonwealth University, Virginia, USA.

***Abstract*—Twitter generates the majority of its revenue from advertising. Third parties usually pay to have their products advertised on Twitter through tweets, accounts and trends. However, spammers can use Sybil accounts (fake accounts) to advertise and avoid paying for it. Sybil accounts are highly active on Twitter performing advertising campaigns to serve their clients. They aggressively try to reach a large audience to maximize their influence. These accounts have similar behavior if controlled by the same master. Most of their spam tweets include a shortened URL to trick users into clicking on it. Also, since they share resources with each other, they tend to tweet similar trending topics to attract a larger audience. However, some Sybil accounts do not spam aggressively to avoid being detected, rendering it difficult for traditional spam detectors to be effective in detecting Sybil accounts with low spamming activities. In this paper, we investigate additional criteria - spam patterns, to measure the similarity across accounts on Twitter. We propose an algorithm to define the correlation among accounts by investigating their tweeting patterns and content. Our real data evaluation reveals that, given known some initially labelled spam tweets, this approach can detect additional spam tweets and spam accounts that are correlated to the initially labelled spam tweets, which are not detected by traditional spam detection approaches otherwise.**

I. Introduction

Online Social Networks (OSNs) have become an essential platform for people in daily communication. The most popular social network websites such as Facebook, Twitter, and Instagram have exceeded billions of users and millions of active users everyday in total. People use these platforms to communicate through messages, posts, links, tweets etc., and share pictures to keep in touch with friends or follow their favorite celebrities. Companies and businesses also use social networks to advertise their products and services, while governmental institutions use them to inform/educate consumers and for emergency situations [1]. Among the most popular social networks, Twitter is designed for users to share news, events, and information by posting messages, links, and pictures. Nowdays the majority of Twitter's revenue is from advertisement. Companies pay to post advertisement on Twitter. Twitter Advertisement reaches an incredible number of users on Twitter and the revenue it produces represents 85% of Twitters overall revenue [2].

However, dishonest companies or attackers may use fake or compromised accounts to advertize by spamming or phishing. The latter can be highly dangerous to legit users, since acquiring sensitive information such as user-names and passwords can be potentially used to access personal information and cause financial loss to the victims. On the other side, not all advertisement on Twitter is spam. Honest companies pay Twitter to endorse their products or services. However, dishonest companies may choose to spam through Twiter to reach their goals with lower cost, and this can be done by purchasing or renting fake accounts to promote their products and services aggressively. For example, 1000 fake accounts only costs $ 11 in the black market [3].

In order to protect their users from being spammed, Twitter has a spam detection systems which can detect suspicious spam accounts, resulting in their suspension [4]. The algorithm focuses on criteria such as harmful links, aggressive following behavior, posting repeatedly to trending topics, posting duplicated tweets, posting links with unrelated tweets etc. However, spam accounts are constantly evolving to avoid being detected [5]. Spam detection systems also have to be improved continuously to remain effective. For example, underground spam account sellers were able to change the behavior of their fake accounts to evade detection by providing more comprehensive account details such as pictures, personal information, and create some activities for the accounts before selling them [6]. If a Twitter spam detection system relies on the criteria of incomplete profiling (no pictures and little activity) to detect fake accounts, then it would be ineffective to detect these new spam accounts.

In this paper we propose a novel approach which does not rely on the individual account profile and activities, but rather on the similarity of spam accounts. For example, the interaction between accounts, how they tweet and re-tweets, and the patterns they use for tweeting. This approach is inspired by the fact that most spam accounts are crafted and controlled by botnets [7], [8]. The spam accounts which are controlled by the same bot master tend to follow similar patterns such as advertising for the same company or product, or using similar patterns for their tweets. We use a bipartisan model to feature the connection between spam and spam accounts, and use an iterative model to compute their likelihood of being spam and spam accounts. With a small number of pre-labeled spam using individual spam labeling criteria, we can find out other hidden spam and spam accounts through the interaction between them. Our evaluation based on real Twitter data confirms that our method is effective in detecting both spam and spam accounts. To the best of our knowledge, this method is the first of the same kind that uses interation between tweets and accounts for spam detection.

II. Related Work

Spam Bots or Sybil accounts trading is very active in the black market. These fake accounts are disguised to be legitimate and distinct users. However, their behaviors are similar [9]. Some of these paid advertisement accounts might seem highly similar to legitimate account, and they cause social network platform millions of dollars in revenue loss each year [2].

There are many different methods proposed in the literature to defend against Sybil accounts. Alvisi, et al. proposed SoK [10] as a new and more accurate method of sybile detection in social networks. In their proposal, they use random walks and white-listing honest nodes to compute the trustworthiness of nodes. They suggest that Sybil defenders should build local defense systems in each community, rather than rely on a global defense system. They argue that such discovery methods that use random walks in local communities are more accurate at discovering Sybil accounts than many other existing methods. Based on the same assumption the authors of [11] illustrate a methodology of nodes labeling in the network as honest user or Sybils.

The author of [12] proposed new criteria of spam accounts which involve publishing embedded links that lead to malicious sites, posting duplicated tweets, sending unsolicited replies and mentions, and using trending topics. All the above criteria focus on individual accounts and do not take into account the relationship between accounts. Spam has evolved by changing their characters in tweets that makes duplicated tweet detection impossible. Tweets need to be stripped of all non-characters, digits and white spaces in order to get a closer match. The work of [13] makes use of a word frequency count by analyzing spam tweets and then query for more tweets that contain these frequent words.

There are also spam accounts that are maintained by humans [14], and they act in a different manner than machine generated accounts. The difference lies in the way the two accounts tweet and maintain their profile. As described in [15], Crowdsourcing Sybil detection is highly efficient in discovering spam accounts when the participants pay close attention to all the accounts they label. However, this way of detection is only applicable on a small scale, as it is time and resource consuming and hence needs to be automated. According to [16], 67% of internet users use social networking sites and 16% use Twitter. Manual discovery of spam accounts is not feasible on such a large scale. While machine maintained spam accounts might have a few random pictures, or simply artistic pictures of nature or sports etc., manually maintained accounts can have many more pictures of a single person, possibly videos and all tweets are very different in wording. In this work we focus on spam accounts which are maintaned by machines which follow some patterns and do not specifically looking for manually maintaned spam accounts.

III. Spam Discovery Method

In this section, we present our method to detect spam tweets as well as distinguish between spammers and legitimate users in Twitter. The proposed method consists of three stages. *First*, we identify Tweets that contain malicious links. For this purpose, we leverage Twitter's database of potentially harmful URLs. However, any malicious links detection methods can be used in this stage. *Secondly*, we extract unique patterns in the spam Tweets and apply a carefully-constructed and conservative pattern matching method to identify additional spamming Tweets. *Finally*, we construct a bipartite network between users and their corresponding tweets and apply an iterative procedure to compute spam scores for both the users and the tweets. The spam scores of tweets indicate the likehood of being spam messages and the users' spam scores indicate the likehood of users being spammers. These scores not only help to identify spammers and spam Tweets, e.g., using simple thresholds, but also give ranking of top/most common spamming users/tweets.

The above stages of our method are detailed in Algorithm 1. The goal of the algorithm is to assign an spam likelihood score to each user and each tweet. Then thresholds can be inferred to give fine lines between spam and non-spam users/tweets.

Algorithm 1 Discovering Spam Accounts and Tweets

Require:

- $U = \{u_1, u_2, ...u_n\}$ set of tweets
- $X = \{x_1, x_2...x_m\}$ set of users

Set $u_j^0 \leftarrow 0 \ \forall j = 1..n$
Stage 1:
for each $u_j \in U$ **do**
  if the jth tweet is flagged by Twitter, set $u_j^0 \leftarrow 1$
end for
Stage 2:
Extract unique patterns from the flagged tweets.
For each unflagged tweet j, if it matches one of the extracted patterns, flagg it and set $u_j^0 \leftarrow 1$
Stage 3:
Set $x_i^0 \leftarrow 0 \ \forall i = 1..m$
$t \leftarrow 0$
repeat
  $t \leftarrow t + 1$
  $x_i^t = \alpha \frac{1}{|N(i)|} \sum_{i \epsilon N(i)} u_j^{t-1} + (1 - \alpha) x_i^{t-1}$
  $u_j^t = \alpha \frac{1}{|N(j)|} \sum_{j \epsilon N(j)} x_i^{t-1} + (1 - \alpha - \beta) u_j^{t-1} + \beta u_j^0$
until $\|U^t - U^{t-1}\| + \|X^t - X^{t-1}\| < \epsilon$
SpamTweets $\leftarrow \{ \ i \ | u_i > \tau \}$
SpamUsers $\leftarrow \{ \ j \ | x_j > \tau \}$
return SpamTweets and SpamUsers

Stage 1: Identify Tweets with Malicious Links. The algorithm starts by assigning a spam score one to all tweets that contain malicious links and values zeros to the rest. To detect Tweets with malicious links, for each account (user), we iterate through each tweet and follow the shortened URL. We then decide whether URL has been flagged based on whether Twitter display warning site.

Twitter has its own mechanism to decide whether a website

TABLE I
SPAM USAGE OF DIFFERENT MENTIONS

Number	Tweet
1	@Lorin_Marie Make An Incredible Income - Follow The Simple Steps http://t.co/NhghOoSJ
2	@lovely_lauren19 Make An Incredible Income - Follow The Simple Steps http://t.co/NpqkGerf
3	@DrTiaCMTyree How to Make Money on the Internet http://t.co/NhghOoSJ
4	@TheOaklandPress How to Make Money on the Internet http://t.co/NhghOoSJ
5	@stargaryen How to Make Money on the Internet http://t.co/Evq7uBT0

is spam or not. We label a website as spam if Twitter or other used URL shortening web service(Google, bit.ly etc.) flagged it as such. A flagged website will prompt the user before displaying the end link. We implemented a web crawler to check if the short URL leads to a warning page. The web crawler looks for a set of predefined warning pages. If a warning page is encountered, the tweet gets marked as spam and assigned an initial spam value of 1, otherwise assigned a value of 0.

Stage 2: Mining Spam Patterns. To discover similar spam tweets, assigning an initial spam value 1 to all tweets with a flagged URL is insufficient. We make use of another method that involves matching tweets by pattern. It starts by creating a hash value for each tweet and comparing it to the hash value of other tweets. The initial analysis of the tweet pool shows that simply hashing the entire tweet is insufficient as spammers tend to alter the tweet by using different characters or digits. Table I illustrates how spam accounts use different mentions to target a particular audience using the same exact tweet.

For this reason, each individual tweet, will be stripped of all non-alpha numberic characters such as digits 0-9 or characters like *,!,@,#. Furthermore any link that starts with http or https will be removed from the tweet, as well as any mentions of other users that start with @user or any hashtags that start with #hashtag. Our pattern extracting technique insures that similar tweeting patterns will be recognized among the tweet pool. Specifically, we match hashes of tweets that have been initially labeled as spam. This method allows the algorithm to discover malicious tweets that might contain different links, different mentions or characters etc, but are at core similar to an already flagged tweet. This ensures the discovery of tweets that have not yet been labeled as spam. It turns out that if a tweet u_j has the same hash as a malicious tweet u_{j+1}, the likelihood of that tweet u_j to be spam will be significantly increased. For this reason We set all initial spam scores of tweets that have the same hash with an initial flagged tweet equal to 1.

Stage 3: Spam Likelihood Estimation. The final step is to estimate the spam likelihood of users and tweets. At this stage, the algorithm analyzes interaction between users and use of similar tweets. When a user tweets multiple spam tweets, the user will also be assigned a high spam likelihood score. The same idea applies to tweets. If a tweet is tweeted by many spam accounts, the tweet will receive a high spam likelihood score.

Using the link crawling and tweet hashing methods described in the above subsections, We devise a mathematical formula that uses the initially assigned spam values of tweets, to assign a final spam value to all tweets and users. It was obvious to me from an initial analysis of the tweets and accounts, that there is a correlation between spam tweets and accounts. Whether this correlation is repeated spam tweets or similar tweeting patterns among spam accounts, it is important to use this correlation to assign a final spam value to tweets and users.

The mathematical formula determines how likely it is for a user or a non-flagged URL to be malicious. It assigns a value between 0 and 1 to each user and tweet. These values are store in the vectors X (representing all the users spam scores) and U (representing all the tweets spam scores). Our updating formulations are based on the following assumptions:

1) The more malicious URLs a user tweets, the more likely it is for the user to be a malicious account.
2) The more a URL is tweeted by malicious users, the more likely it is for the URL to be spam.

Initially all elements of the set X are equal to 0, which means that all users are considered to be non-malicious. The spam score of a user in a round depends on the spam scores of URLs that user tweeted as well as the spam score of that user in the previous round. The constant α decides how much of the score in the previous round we retain. The constant β is to force all tweets with marked URLs to have a minimum spam score. At the end of the algorithm, each element in this set will be equal to a real number $x_i \epsilon$ [0,1] which represents the likelihood of being a malicious user. On the other hand, all elements in the set U will initially be either 0 or 1. The elements in this set that have been set to equal 1, are the URLs that have already been flagged as malicious by Twitters defense system or by the hashing method described in the previous section. All other URLs in the set U associated with the value 0, have not yet been classified as being malicious or truthful. Having a value of 0, does not equate to being truthful, it means it has yet to be classified.

The first step presented in the following formula, assigns the initial spam value to all users in X to be zero and all tweets in U are assigned values according to the first two stages. At an iteration $t \geq 1$, users and tweets spam scores will be updated, based on the interaction of the malicious users with non-malicious users and common tweets among them. That is the scores are updated using the following recursions.

$$x_i^t = \alpha \frac{1}{|N(i)|} \sum_{i \epsilon N(i)} u_j^{t-1} + (1-\alpha) x_i^{t-1} \quad (1)$$

$$u_j^t = \alpha \frac{1}{|N(j)|} \sum_{j \epsilon N(j)} x_i^{t-1} + (1-\alpha-\beta) u_j^{t-1} + \beta u_j^0 \quad (2)$$

βu_j^0 used in order to assign a higher value to a URL that was initially marked as spam. The variable α decides how much of the spam score in the previous round influence the current spam score of a user and β is the minimum score enforced for all tweets that are marked as spams.

The computation converges when

$$\|U^t - U^{t-1}\| + \|X^t - X^{t-1}\| < \epsilon \quad (3)$$

where $\epsilon > 0$ is a predefined threshold ($\epsilon = 0.001$ in our experiments).

Our experiments indciate that the proposed method is robust under different values of α and β. Changing values of α and β only affect the number of iterations that the algorithm takes to converge but not the likehood scores. Moreover, the spam threshold τ is fixed to be 0.1. Even choosing a threshold as high as $\tau = 0.3$ has insignificant effect on the classifaction of spamming users and tweets.

IV. Evaluation and Results

In this section, we present the evaluation of our proposed spam detection method through real data from Twitter. We first present data collection and then some preliminary analysis on collected data. We then evaluate our spam detection algorithm.

A. Data Collection

To evaluate the effectiveness of our spam detection algorithm, we collect real Twitter data from Twitter website. We chose to Collect new data from Twitter because all the public available Twitter data sets available are outdated. Spam evolves constantly to avoid detection. Therefore it is necessary to have new data for evaluation.

The data collection algorithm started with a random Twitter account from which we further downloaded a list of followers. we then randomly selected 200 followers from that list. We repeated the process until we obtained sufficient number of accounts. Previous research showed that Amazon and Toyota were running major advertisement campaigns on Twitter towards the end of 2013. therefore we downloaded random users from the list of followers of each of these companies.

We collected data over the course of 6 weeks using the Java API provided by Twitter for the data collection and stored them in a mySQL database. we were able to download approximately 10 Million tweets from 51,000 user accounts. For each user we recorded the basic profile and the number of followers and friends, etc.

B. Experiment Setup

The experiments were conducted by running bash shell scripts on a linux server. To obtains some initial spam tweets to bootstrap the algorithm, we label all tweets which contain malicious URLs to be spam. After pattern matching process, we were able to label many more tweets to be spam. At the beginning all unlabelled tweets and accounts have initial score 0. Labelled spam tweets have score 1. In the following subsections we present the evaluation results of our algorithm.

TABLE III
Statistics

	μ	$\widetilde{u}$	max	min	σ
All tweets	0.00056	4.5e-09	0.99999	0	0.01788
spam=1	0.73564	0.66417	0.99999	0.625	0.13513
spam=0	0.00015	4.5e-09	0.36499	0	0.00244
All users	0.00056	1.17e-08	0.99999	0	0.01284

C. Algorithm Convergence

There are several parameters, such as α, β and ϵ are used in the algorithm. In this experiment we are interested to know the impact of those parameters to the convergence speed of the algorithm. Table III shows the impact of different values of α, β and ϵ to the number of iterations the algorithm takes to converge, the number of users assigned a spam score higher than 0.1 and the X and U vector distance (from X and U when $\alpha = 0.1$ and $\beta = 0.2$):

TABLE II
Vector Distance and Iterations

α	β	Iterations	Users>0.1	X diff	U diff
0.1	0.1	217	108	1.42E-5	1.82E-4
0.1	0.2	144	109	-	-
0.1	0.5	101	109	9.23E-6	1.81E-4
0.1	0.8	89	109	1.16E-5	2.42E-4
0.2	0.1	194	107	3.97E-5	3.66E-4
0.2	0.2	116	108	1.42E-5	1.82E-4
0.2	0.5	68	109	3.10E-6	5.20E-5
0.5	0.1	181	104	1.00E-4	5.56E-4
0.5	0.2	101	104	5.13E-5	4.19E-4
0.8	0.1	178	100	1.48E-4	6.22E-4

We can see that α and β affect the convergence of the algorithm. A smaller α and a smaller β results in a higher number of iterations. When α increases from 0.1 to 0.2 and β increases from 0.2 to 0.5 the number of iterations decreases significantly. Another observation is the fact that the spam values assigned to a tweet or an account is either much higher(if it is spam) or much lower(if not spam) when α and β are smaller. This is due to the fact that the number of interaction increase as α and β decrease.

We can see that the vector distance of both U and X approach to 0 exponentially, which means an abrupt decline in the distance at the beginning and slow convergence after the initial drop.

D. Converged Spam Likelihood Results

Table IV shows running results that all tweets initially flagged as spam have a spam likelihood value of 0.625 (or higher) whereas the maximum value is 0.9999976. The average spam likelihood of these tweets is μ=0.7356384 and the median is 0.6641736.

For tweets that are not initially labelled as spam by Twitter start with a initial spam value of 0. At the end of the iterations the algorithm has assigned the maximum spam value of 0.3649912 and the minimum value of 0. The mean of these values is μ=0.0001477763 while the median is 4.5e-09.

An analysis of the tweets database shows that 4262 tweets have been labeled as spam with a spam value greater than 0.1. Out of these 4262, 1991 were initially labeled as spam and the rest had an initial spam score=0;

To verify the accuracy of the spam detection algorithm We manually evaluated the highest ranked spam accounts (initially not flagged by Twitter). To verify whether a tweet is indeed spam or not, We checked the URL embedded in the tweet. We found the following results:

1) Highest ranked tweets: Out of 200 highest ranked tweets, we found that:

- All tweets were spam.
- 172 URLs were either blocked or dead.
- 28 URLs were accessible.

By analyzing the URL's that were accessible(28), We found that all of them can be categorized as spam. They represent different websites that appear to be aggressive advertisement, with links to a myriad of online social websites, peer-to-peer websites and different product endorsement.

2) Highest ranked tweets without initial labelling: Out of tweets with the highest scores that were not initially ranked as spam, we found that:

- A tweet composed of a link only to the highly popular(440M+ views) Youtube video of the Original Gymmy Bear song
- The 2nd highest is a tweet composed on only a link that has been blocked by Google
- Tweet: "SenFeinstein as one of your constituents, We ask that you support H.R.6480 and S.3609 IRFA: http://t.co/ndOo3x8l #FairNetRadio"

It turns out that the highly popular video link, of the Youtube video of the Original Gummy Bear song, was a false positive. Since spamming accounts may also tweet popular topics, hashtags and links to try and reach a broader audience, false positives can also appear.

3) Highest ranked tweets with initial labelling: Out of tweets with the highest scores that were initially marked as spam, we found the following examples:

- check this out! We made almost $600 today so far http://t.co/m4e1PvU
- hey everyone youve got to check this out We made almost $500 today! http://t.co/ZURqqrk
- Walk out of your crappy 9-5 job this week! http://t.co/1XjQUr9t
- http://t.co/CeWjdnTcVW We could make some serious money selling nude pics of myself to bulimics with short fingers.
- I'm gonna start my own TV network called RealityTV(RTV) and play nothing but music videos http://t.co/R2CBvXXyIZ

As a conclusion, our algorithm discovered additional 2271 tweets, that were not initially labelled as spam. These tweets are associated with 1830 distinct users.

E. Score Distribution of Spam Tweets

Figure 2 shows the distribution of the tweets with a spam score 0.1(inclusive) or higher. We can see that about 1500 of the total number of these tweets have a final spam score of about 0.2. Also, there is a slow increase trend between 0.65 and 0.85, then about 400 of them with a score very close to 1. This similar spike can also be observed in Figure 3. This is an indication that the algorithm tends to assign either a low score or high score to tweets. Not many tweets can be observed in the range [0.25-0.6] regardless of α and β. Most users that tweet spam tweets, either tweet very few or many spam tweets. The users that tweet very few spam tweets, are likely be compromised accounts for a short period of time. However most of the spam tweets that they tweet, are also tweeted by spam accounts. If a tweet is tweeted by a spam account and a honest account at the same time, the tweet will be assigned a high spam score. If a tweet is tweeted by only an honest account (that tweets a very small spam to non-spam ratio), but not any spam account, the tweet will finally be assigned a spam score below the 0.1 threshold. However, if a tweet is tweeted by many spam accounts,it will be assigned a high spam score, above 0.6 regardless whether an honest user also tweeted it or not.

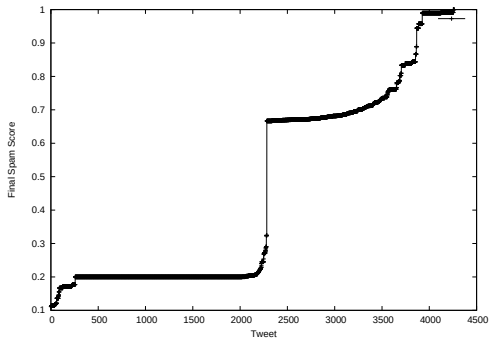


Fig. 1. Tweets Spam Distribution

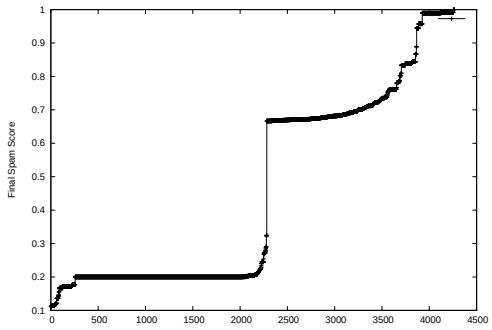
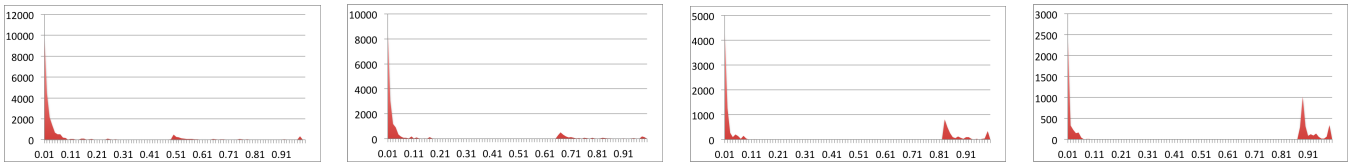


Fig. 2. Initial Spam Tweet Final Spam Score Distribution

Analyzing the users that have been ranked with a score higher than 0.1, 108 of them were finally marked as spam. Manual analysis of the 108 highest ranked accounts shows that 102 are spam accounts while 6 seem to non-spam, legit users. On the other side, it is interesting to observed that all the initial spam tweets that were labeled as malicious by Twitter (or other URL shortening tools) have received a final score of 0.65 or higher. This distribution of spam scores for initially flagged tweets can be seen below: All initially flagged tweets received a final spam score of 0.65 and higher.

Figure 4 illustrates the distribution of the spam score vector U (for Twitter accounts) during different runs.



(a) $\alpha = 0.1, \beta = 0.1$ (b) $\alpha = 0.1, \beta = 0.2$ (c) $\alpha = 0.1, \beta = 0.5$ (d) $\alpha = 0.1, \beta = 0.8$

Fig. 3. Tweets spam score distributions

The vector U contains 3.5M tweets and each index in the vector represents the spam score (for a tweet). When β is large most of the spam scores are within the [0.0-0.1] and [0.8-0.99] range. This is because in the mathematical equation used to discover spam, βu_j^0 will keep the initial spam score of the tweet. From these graphs, the following assertions can be made:

- A tweet with a score below 0.1 is not spam
- A tweet with a score above 0.8 is highly likely spam

F. Sample Results

Given an initial set of spamming tweets, users that have tweeted multiple flagged tweets, have resulted with a high spam likelihood score. The majority of these tweets are spam, and easy to spot. It is not surprising that the following tweets have received a spam score as high as 0.9999:

- @mildsto*** Real ways to make money using computers and the Internet http://t.co/NhghOoSJ
- want to start your own business in 2013? look at this - http://t.co/kJIAtUjm

These tweets are obvious spam and expected to have a high spam score. Similar tweets that were initially not flagged but have been tweeted by many spam accounts are also expected to have a high spam score. The following tweets are examples of obvious spam, which were not flagged by Twitter, but were discovered as spam:

- My best week! Earned $231.35 doing surveys in past week :) LOOK http://t.co/f6dTPIFk
- Just downloaded the Webs Best Investment Sites magazine by Old School Value Jae_Jun http://t.co/zx158HTHMl

By analyzing tweets that received a high spam score, there are certain patterns that can be observed. Most of these tweets give an incentive to the user to click on the link. It is interesting to see the similar pattern/phrasing they use. A popular trending topic observed is dietary advice, mostly the promise to lose weight by using a certain product or following a diet or promotion of certain weight loss product such as raspberry ketones or green coffee beans.

V. Conclusion

Traditional Twitter Spam detectors focus on the spamming behaviors and can not detect spam accounts which do not spam aggressively in social networks. Therefore, they are not effective to detect less active spam accounts and their spam tweets. In this paper we presented a novel Twitter spam discovering method, by analyzing the relationship between accounts based on their tweeting pattern similarity. An spam score computation algorithm is proposed to iteratively update the spam scores of users and tweets based on their pattern similarity and their closeness to known initially labelled spam tweets. Our experiment based on real data demonstrate that a substantial amount of new spam tweets and spam accounts are discovered by our proposed method which are otherwise not detected.

References

[1] R. Beneito-Montagut, S. Anson, D. Shaw, and C. Brewster, "Governmental social media use for emergency communication," in *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management, Baden-Baden, Germany*, 2013.

[2] I. A. f. w. t. S. Twitter and . Exchange Commission on October 15, "Twitter inc. s1 form," 2013.

[3] J. D. B. Labs, "Twitter underground economy still going strong," 2013, http://www.net-security.org/article.php?id=1859&p=1.

[4] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: An analysis of twitter spam," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '11. New York, NY, USA: ACM, 2011, pp. 243–258.

[5] D. Wang, D. Irani, and C. Pu, "A perspective of evolution after five years: A large-scale study of web spam evolution," *International Journal of Cooperative Information Systems*, 2014.

[6] J. Elder, "Inside a twitter robot factory," The Wall Street Journal, 2013.

[7] I. Adegbola, R. Jimoh, and O. Longe, "An integrated system for detection and identification of spambot with action session and length frequency," *Computing, Information Systems, Development Informatics and Allied Research Journal*, vol. 4, no. 2, 2013.

[8] I. Adegbola and R. Jimoh, "Spambot detection: A review of techniques and trends," *network*, vol. 6, no. 9, 2014.

[9] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "Sybilguard: Defending against sybil attacks via social networks," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 4, pp. 267–278, Aug. 2006.

[10] L. Alvisi, A. Clement, A. Epasto, S. Lattanzi, and A. Panconesi, "Sok: The evolution of sybil defense via social networks," *2012 IEEE Symposium on Security and Privacy*, vol. 0, pp. 382–396, 2013.

[11] G. Danezis and P. Mittal, "Sybilinfer: Detecting sybil nodes using social networks," 2009.

[12] A. H. Wang, "Don't follow me: Spam detection in twitter," in *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*, July 2010, pp. 1–10.

[13] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th Annual Computer Security Applications Conference*, ser. ACSAC '10. New York, NY, USA: ACM, 2010, pp. 1–9.

[14] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson, "Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse," in *Proceedings of the 22Nd USENIX Conference on Security*, ser. SEC'13. Berkeley, CA, USA: USENIX Association, 2013, pp. 195–210.

[15] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. J. Metzger, H. Zheng, and B. Y. Zhao, "Social turing tests: Crowdsourcing sybil detection," *CoRR*, vol. abs/1205.3856, 2012.

[16] B. J. Duggan M., "The demographics of social media users âĂŤ 2012," 2013.