Twitter: Who gets Caught?

Observed Trends in Social Micro-blogging Spam

Abdullah Almaatouq
Center for Complex Engineering Systems at KACST and MIT
amaatouq@mit.edu

Ahmad Alabdulkareem
Center for Complex Engineering Systems at KACST and MIT
kareem@mit.edu

Mariam Nouh
Center for Complex Engineering Systems at KACST and MIT
mnouh@kacst.edu.sa

Erez Shmueli
Massachusetts Institute of Technology (MIT) Media Lab
shmueli@mit.edu

Mansour Alsaleh
King Abdulaziz City for Science and Technology
maalsaleh@kacst.edu.sa

Vivek K. Singh
Massachusetts Institute of Technology (MIT) Media Lab
singhv@mit.edu

ABSTRACT

Spam in Online Social Networks (OSNs) is a systemic problem that imposes a threat to these services in terms of undermining their value to advertisers and potential investors, as well as negatively affecting users' engagement. In this work, we present a unique analysis of spam accounts in OSNs viewed through the lens of their behavioral characteristics (i.e., profile properties and social interactions). Our analysis includes over 100 million tweets collected over the course of one month, generated by approximately 30 million distinct user accounts, of which over 7% are suspended or removed due to abusive behaviors and other violations. We show that there exist two behaviorally distinct categories of twitter spammers and that they employ different spamming strategies. The users in these two categories demonstrate different individual properties as well as social interaction patterns. As the Twitter spammers continuously keep creating newer accounts upon being caught, a behavioral understanding of their spamming behavior will be vital in the design of future social media defense mechanisms.

Categories and Subject Descriptors

H.0 [**Information systems**]: General; K.4.2 [**Social issues**]: Abuse and crime involving computers; H.2.8 [**Database Applications**]: Data mining

Keywords

Spam; Online Social Networks; Microblogging; Account Abuse

1. INTRODUCTION

Spam exists across many types of electronic communication platforms, including email, web discussion forums, text messages (SMS), and social media. Today, as social media continues to grow in popularity, spammers are increasingly abusing such media for spamming purposes. According to a recent study [21], there was a 355% growth in social spam during the first half of 2013. Twitter company's initial public offering (IPO) filing indicates spam as a major threat in terms of undermining their value to advertisers and potential investors, as well as negatively affecting users' engagement [32].

While there is a growing literature on social media in terms of developing tools for spam detection (e.g., [18, 24, 33]) and analyzing spam trends (e.g., [27, 37, 38]), spammers continue to evolve and change their penetration techniques. Therefore, there is a continuous need for understanding the evolving and diverse properties of malicious accounts in order to combat them properly [21, 32].

In this paper, we present an empirical analysis of spam accounts on Twitter, in terms of profile properties and social interactions. The analysis includes identifying categories (sub-populations) of spam accounts (see Section 4). Through profile analysis we identify distinct characteristics and patterns that pertain to different identified categories of Twitter accounts (see Section 5). We also examine the network properties of several social interactions (namely, follow relationship and mention) to improve our understanding of the methods used by spammers for reaching spam victims (see Section 6).

To perform the study, we collected over 100 million tweets over the course of one month (from March 5, 2013 to April 2, 2013) generated by approximately 30 million distinct user accounts (see Section 3). In total, over 7% of our dataset accounts are suspended or removed accounts due in part to abusive behaviors and other violations. The summary and future work of our study discussed in Section 8.

In summary, we frame our contributions as follows:

- We categorize spam accounts based on their behavioral activities and find that Twitter spammers belong to two broad behavioral categories. We observe that these categories of spam accounts exhibit different spamming patterns and employ distinct strategies for reaching their victims, and should therefore be analyzed separately and treated differently by future social media defense mechanisms.

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.

http://dx.doi.org/10.1145/2615569.2615688.

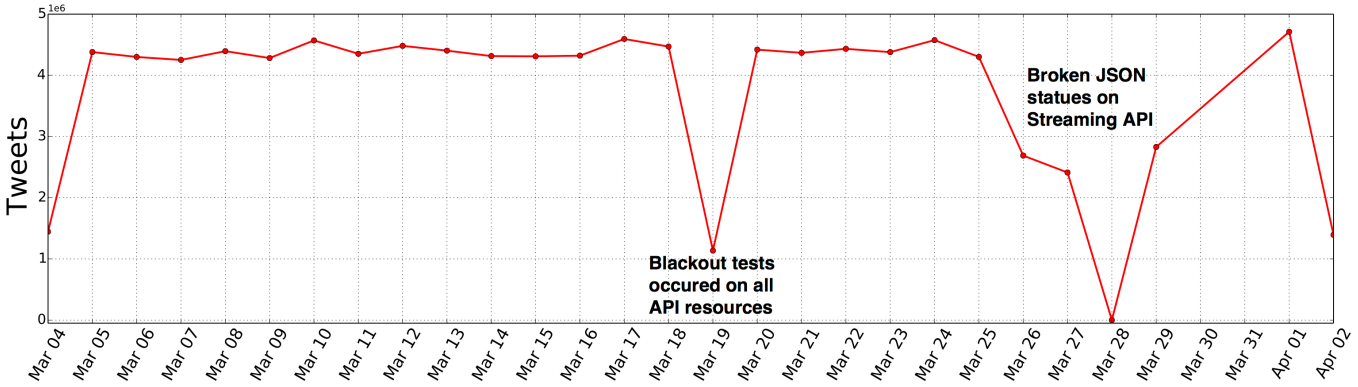


Figure 1: Tweets received per day. On average, we receive 4 million tweets per day

- We analyze the different properties of spam accounts in terms of their profile attributes and use the attributes of legitimate accounts as a baseline. From this, we identify a cluster of malicious accounts that seems to be originally created and customized by legitimate users, whereas the other cluster deviates from the baseline significantly.

- Through network analysis of multiple social interactions, we reveal a set of diverse strategies employed by spammers for reaching audiences. We focus on the mention function as it is one of the most common ways in which spammers engage with users, bypassing any requirement of sharing a social connection (i.e., follow/following relationship) with a victim.

2. BACKGROUND

Twitter is a micro-blogging platform and an Online Social Network (OSN), where users are able to send *tweets* (i.e., short text messages limited to 140 characters). According to a recent study, Twitter is the fastest growing social platform in the world [12]. In 2013, Twitter estimated the number of active users at over 200 million, generating 500 million tweets per day [32].

Twitter spam is a systemic problem [27]. While traditional email spam usually consists of spreading bulks of unsolicited messages to numerous recipients, spam on Twitter does not necessarily comply to the volume constraint, as a single spam message on Twitter is capable of propagating through social interaction functions and reach a wide audience. In addition, previous studies showed that the largest suspended Twitter accounts campaigns directed users via affiliate links to some reputable websites that generate income on a purchase, such as Amazon [27]. Such findings blur the line about what constitutes as OSN spam. According to the "Twitter Rules", what constitutes *spamming* will evolve as a response to new tactics employed by spammers [31]. Some of the suspicious activities that Twitter considers as indications for spam [31] include: (1) aggressive friending; (2) creating false or misleading content; (3) spreading malicious links; and (4) trading followers.

Spam content can reach legitimate users through the following functions: i) *home timeline*: a stream showing all tweets from those being followed by the user or posts that contain *@mention* requiring no prior follow relationship; ii) *search timeline*: a stream of messages that matches a search query; iii) *hashtags*: tags used to mark tweets with keywords or topics by incorporating the symbol # prior to the relevant phrase (very popular hashtags are called *trending topics*); iv) *profile bio*: spam accounts generate large amounts of relationships and favorite random tweets from legitimate users with the hope that victims would view the spammer account profile which often contains a URL embedded in its bio or description; and v) *direct messages*: private tweets that are sent between two users.

Accounts distributing spam are usually in the form of: i) *fraudulent accounts* that are created solely for the purpose of sending spam; ii) *compromised accounts* created by legitimate users whose credentials have been stolen by spammers; and iii) legitimate users posting spam content. While, multiple previous studies focused on fraudulent accounts(e.g., [27, 28]), the compromised accounts are more valuable to spammers as they are relatively harder to detect due to their associated history and network relationships. On the other hand, fraudulent accounts exhibit a higher anomalous behavior at the account level, and hence are easier for detection [9].

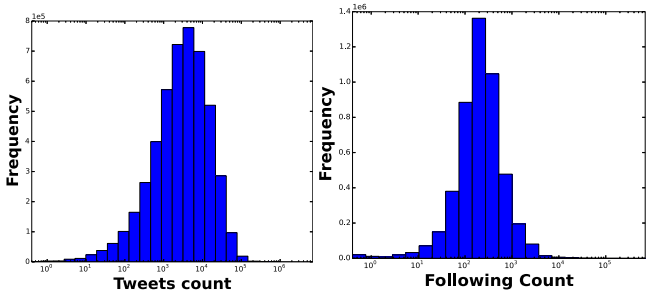
3. DATASETS

Our Twitter dataset consists of 113,609,247 tweets, generated by 30,391,083 distinct users, collected during a one month period from March 5th, 2013 to April 2nd, 2013 using the Twitter public stream APIs [30]. For each tweet, we retrieve its associated attributes (e.g., tweet text, creation date, client used, etc.) as well as information tied to the account who posted the tweet (e.g., the account's number of following, followers, date created, etc.). On average, we receive over 4 million tweets per day. We lack data for some days due to network outages, updates to Twitter's API, and instability of the collection infrastructure (using Amazon EC2 instances). A summary of tweets collected each day and outage periods is shown in Figure 1.

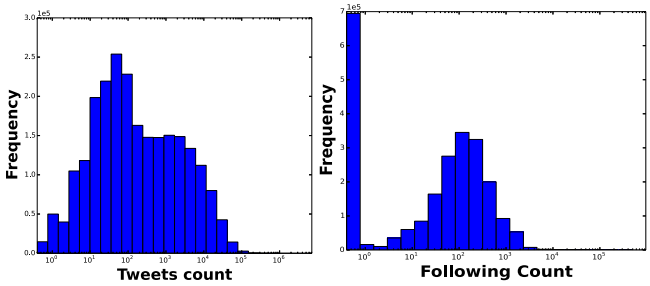
In order to label spammer accounts in our dataset, we rely on Twitter's account suspension algorithm described in [27]. Given that the implementation of the suspension algorithm is not publicly available, we verify whether an account has been flagged as spam by checking the user's profile page. In case an account has been suspended or removed, the crawler request will be redirected to a page describing the user statues (i.e., suspended or does not exist). While all of the removed/suspended user's information is no longer available through the Twitter's API, we were able to reconstruct their information based on the collected sample. In total, over 7% of our dataset are suspended/removed accounts. Although the primary cause for suspension or deletion of Twitter accounts is spam-activity, Twitter's policy page states that other activities such as publishing malicious links, selling usernames and using obscene or pornographic images may also result in suspension or deletion [31]. Removed accounts may include users that deactivated their accounts during the data collection period.

4. IDENTIFYING SUB-POPULATIONS

The results of the initial analysis to compare the collective tweeting patterns and social behavior of normal and malicious users showed tendency for bi-modality in the case of spam accounts. This was less evident in the case of legitimate users (see Figure 2). This pattern occurs across multiple attributes (i.e., tweets count, favorites count, followers count, etc.). The bi-modal distributions commonly arises as a mixture of uni-modal distributions corresponding to mixture of populations. Accordingly, we separated the sub-populations within spammers, using Gaussian Mixture Models (GMM), in order to reveal distinct spamming strategies and behaviors.



(a) Non-Spam accounts



(b) Spam accounts

Figure 2: An illustration of different tweeting patterns and following behaviors for normal and spam accounts.

In order to identify subsets of malicious accounts, we use Gaussian Mixture Models (GMM). GMM is a probabilistic model that assumes that data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. To determine the number of components (i.e., sub-populations or clusters) we fit multiple GMMs with different numbers of Gaussians and then calculate the Bayesian Information Criteria (BIC) score for each fit. The use of BIC penalizes models in terms of the number of parameters or complexity. Hence, complex models (i.e., high number of free parameters) will have to compensate with how well they describe the data. This can be denoted as follows:

$$BIC(M_c) = -2 \cdot \ln P(x|M_c) + \ln N \cdot k \qquad (1)$$

where x is the observed data, N is the number of observations, k is the number of free parameters to be estimated and $P(x|M_c)$ is the marginal likelihood of the observed data given the model M with c number of components.

A GMM with two components and spherical covariance gives the lowest BIC score (see Figure 3). The results of the clustering exhibit two classes of spam accounts $C_1 \subset C$ and $C_2 \subset C$, where C is the set of all accounts. We refer to the normal class (i.e., legitimate accounts) as C_{normal}. The results of the separation in one dimension (i.e., tweets count) is shown in Figure 4.

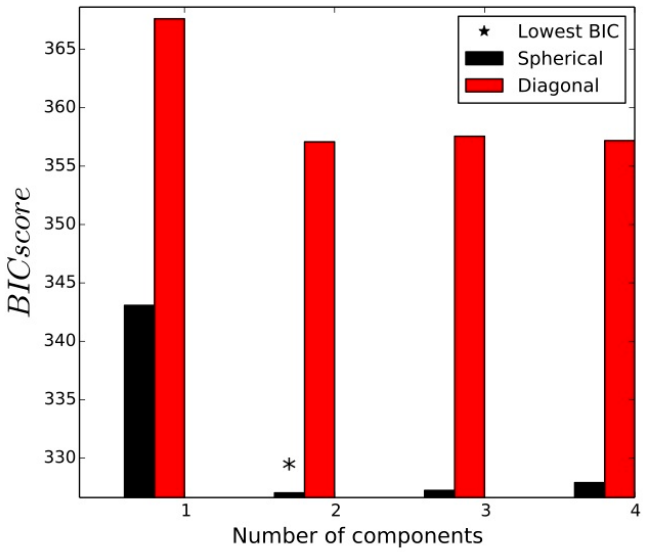


Figure 3: BIC scores for different numbers of components & covariance constraints

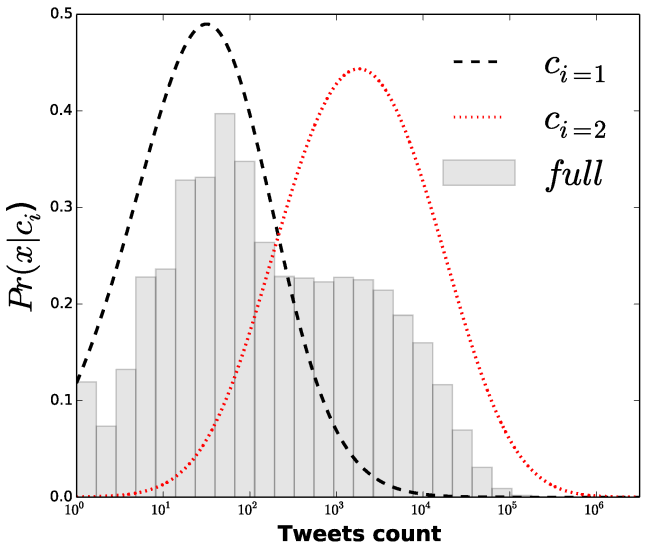
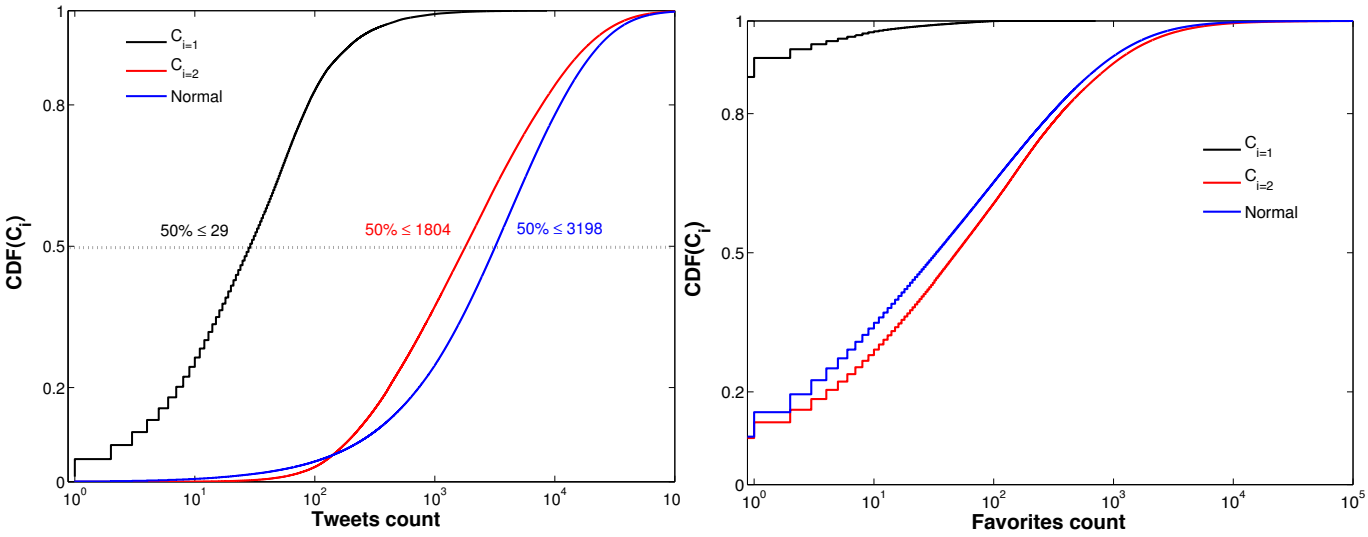


Figure 4: The identified clusters in 1-d (tweets count) for the spam accounts

Based on the separation, we can further investigate the properties and activity patterns of the different identified classes. This separation aids in developing taxonomies and exploit meaningful structures within the spam accounts communities.

5. PROFILE PROPERTIES

In order to further investigate the different identified classes, we examine the Empirical Cumulative Distribution Functions (ECDF) of different attributes for each class (see Figure 5). We find that 50% of the accounts in C_1 have less than 29 tweets, however, for C_{normal} and C_2, 50% of the accounts have tweeted around 2000 times. Furthermore, we find that almost 90% of the accounts in C_1 have no favorites (i.e., tweets added to their favorites list), whereas



(a) Tweets count Empirical Cumulative Distribution Function

(b) Favorites count Empirical Cumulative Distribution Function

Figure 5: Comparison between the three classes C_1, C_2 and C_{normal} in terms of tweeting and following behaviors after the GMM clustering

C_2 and C_{normal} show closely matching patterns, with 50% of the accounts having less than 50 favorite tweets.

We continue to observe similar patterns across multiple attributes, where C_2 and C_{normal} have similar distributions and C_1 deviates from the baseline. We explain this observation through the hypothesis that C_2 mainly consists of *compromised accounts*, while C_1 consists of *fraudulent accounts* as defined in Section 2.

Table 1: Summary of basic profile attributes

	Default profile	Default image	URL	Bio
C_{normal}	22%	1.3%	29%	83.6%
C_1	76%	14%	4%	60%
C_2	36%	1.5%	20%	84.7%

The similarity between C_{normal} and C_2 in the basic profile attributes, such as the percentage of accounts with default profile settings, default profile images, profile descriptions and profile URLs (see Table 1) might indicate that C_2 accounts were originally created and customized by *legitimate users*. For example, we notice that only 22% of C_{normal} and 36% of C_2 accounts kept their default profile settings unchanged, in comparison to 76% in the case of C_1.

6. SOCIAL INTERACTIONS

In this section we analyze users behavior in terms of the follow relationship and mention functions, from the topological point of view. We approach this by incorporating multiple measures that are known to signify network characteristics (differences and similarity). Through this analysis, we reveal sets of behavioral properties and diverse strategies employed by spammers for engaging with victims and reaching audiences.

6.1 Preliminaries

Let $G = (V,E)$ be the graph that represents the topological structure of a given function (i.e., follow or mention), where V is the set of nodes and E is the set of edges. An edge in the graph is denoted by $e = (v,u) \in E$ where $v,u \in V$. Note that in the follow and mention networks, a node v corresponds to a Twitter user and an edge corresponds to an interaction between a pair of users. If two nodes have an edge between them, they are adjacent and we refer to them as neighbors.

We define the neighborhood of node v as the sub-graph $H = (V',E') \mid V' \subset V$ and $E' \subset E$ that consists of all the nodes adjacent to v (alters) excluding v (we refer to v as ego) and all the edges connecting two such nodes. The 1.5 egocentric network $E_{1.5}(v)$ of node v is defined as the neighborhood sub-graph including v itself. Therefore, the neighborhood can be denoted as $N(v) := \{u \mid (u,v) \in E$ or $(v,u) \in E\}$ and the 1.5 ego network as $E_{1.5}(v) := \{N(v) \cup \{v\}\}$.

Focusing on the egocentric networks around the nodes allows for studying the local graphical structure of a given user and signifies the types of interactions that develop within their social partners. Figure 6 shows an illustration of different levels of egocentric networks. From this we can define node properties and measure the relative importance of a node within its egocentric network such as node degree $d(v)$, node out-degree $d_{out}(v)$, in-degree $d_{in}(v)$, and reciprocal relationship $d_{bi}(v)$.

$$\begin{aligned} d_{out}(v) &= |\{u \mid (v,u) \in E_{1.5}(v)\}| \\ d_{in}(v) &= |\{u \mid (u,v) \in E_{1.5}(v)\}| \\ d(v) &= d_{in} + d_{out} \\ d_{bi}(v) &= |\{u \mid (u,v) \in E_{1.5}(v) \wedge (v,u) \in E_{1.5}(v)\}| \end{aligned} \tag{2}$$

From the properties defined in equation 2 we can derive the in-degree density $density_{in}(v)$, out-degree density $density_{out}(v)$, and the density of reciprocal relationships $density_{bi}(v)$.

$$\begin{aligned} density_{in}(v) &= \frac{d_{in}(v)}{d(v)} \\ density_{out}(v) &= \frac{d_{out}(v)}{d(v)} \\ density_{bi}(v) &= \frac{d_{bi}(v)}{d(v)} \end{aligned} \tag{3}$$

In addition, we calculate the betweenness centrality for each ego node in order to quantify the control of such node on the communi-

cation between other nodes in the social network [10]. The measure computes the fraction of the shortest paths that pass through the node in a question v within its egocentric network $E_{1.5}(v)$. Therefore, the betweenness centrality $C_B(v)$ can be computed as [5]:

$$C_B(v) = \sum_{u \neq w \in N(v)} \frac{\sigma_{uw}(v)}{\sigma_{uw}} \quad (4)$$

where σ_{uw} is the total number of shortest paths from node u to node w and $\sigma_{uw}(v)$ is the number of those paths that pass through the node v. Therefore, $C_B(v) = 0$ in the case where all the alters are directly connected to each other and $C_B(v) = 1$ when the alters are only connected to each other through the ego node.

We also compute the closeness centrality $C_C(v)$ which measures the inverse of the sum of the shortest path distances between a node v and all other nodes $u_0, u_1, .., u_n \in N(v)$ normalized by the sum of minimum possible distances. This can be formulated as follows:

$$C_C(v) = \frac{n-1}{\sum_{u \in N(v)} \sigma(v,u)} \quad (5)$$

where $\sigma(u,v)$ is the shortest path distance between v and u, and n is the number of nodes in the egocentric graph.

A network is strongly connected if there is a path between every node to every other node in a directed graph. We define the number of strongly connected components in the egocentric networks $E_{1.5}(v)$ and open neighborhood $N(v)$ to be $SCC_{E_{1.5}}(v)$ and $SCC_N(v)$ respectively. By replacing all of the directed edges with undirected edges, we compute the number of weakly connected components for the egocentric network and open neighborhood as $WCC_{E_{1.5}}(v)$ and $WCC_N(v)$ respectively. The SCC and WCC are used to measure the connectivity of a graph.

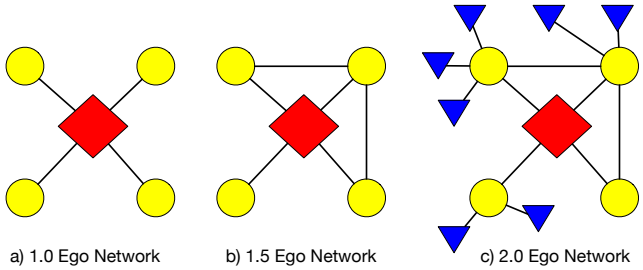
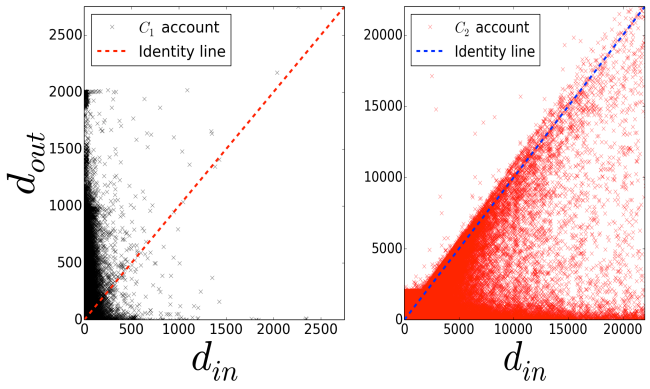


Figure 6: An illustration of the a) 1.0 egocentric network; b) the 1.5 egocentric network; and c) the 2.0 egocentric network. The Ego node is marked in red (diamond) and its connections (alters) are marked in yellow (circles) and the alters' connections are marked in blue (triangles).

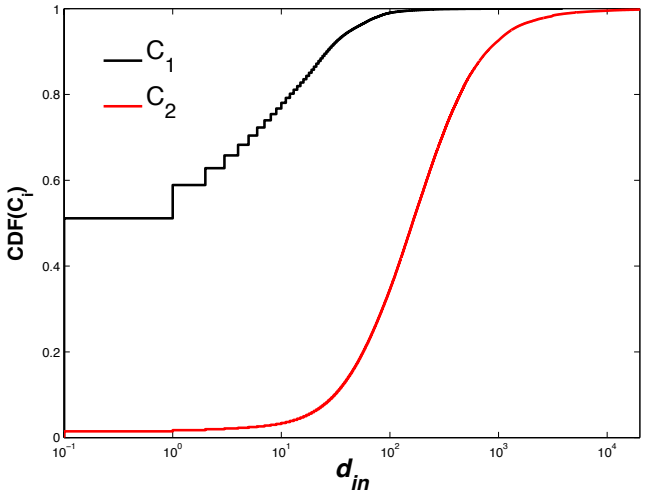
6.2 Relationship Graph

Twitter follow relationship is modeled as a directed graph, where an edge between two nodes $e = (v,u) \in E$ means that v is following u. For the follow relationship, we only have the number of followers and following for each account, and not the actual relationship list. Therefore, in order to compare relationships formed by both C_1 and C_2, we aggregate following and follower data from both classes.

Figure 7 shows the number of followers and following represented by the in-degree d_{in} (follower) and out-degree d_{out} (following) for each class. We find that spam accounts that belong to C_1



(a) Followers d_{in} vs. following d_{out} for C_1 and C_2 accounts.



(b) Followers count Empirical Cumulative Distribution Function.

Figure 7: Illustration of the different relationship behaviors for C_1 and C_2. We find that spam accounts that belong to C_1 are heavily skewed towards following rather than followers or the identity line. The effect of the number of following limit (i.e., 2000 d_{out}) is apparent/observed in both classes.

are heavily skewed towards following rather than followers, which could indicate a difficulty in forming reciprocal relationships. Furthermore, we observe a low $density_{in}$ for C_1 with an average of 0.16 and high $density_{out}$ with an average of 0.4. On the other hand, C_2 has more balanced densities with approximately 0.5 for both.

While Twitter does not constrain the number of followers a user could have, the number of following (i.e., d_{out}) is limited [29]. Every user is allowed to follow 2000 accounts in total; once an account reaches this limit, they require more followers in order to follow more users [29]. This limit is based on the followers to following ratio.

Furthermore, as shown in Figure 7b, almost 50% of C_1 accounts have no followers (i.e., they did not embed themselves within the social graph) and almost 75% of these accounts have less than ten followers. We find that C_2 accounts are more connected in terms of social relationships, which makes them harder to detect and hence contribute more content. These findings adhere to a known phenomenon observed in multiple security contexts. For example, Altshuler et al. [2] showed that in many cases (especially in social networks), optimal attack strategies (i.e., causing greater damage

or spreading more spam content) exhibit slow spreading patterns rather than spreading aggressively.

Table 2: Market prices for followers

Provider	$-per-follower
Socialkik	$0.024
BuyTwitterFriends	$0.003
UnlimitedTwitterFollowers	$0.02

The compromised account population that exists within C_2 can utilize the associated history and network relationships of the original account owner to aid them in increasing the visibility of their spam content. It is also possible that fraudulent and compromised accounts can gain more followers by purchasing them from online services (see Table 2 for recent market prices) to evade detection [36, 37].

6.3 Mention Graph

The mention function is one of the most common ways in which spammers engage with users, unlike the *Direct Messages (DM)*, it bypasses any requirement of prior social connection with a victim.

The mention network is constructed as a simple, weighted, and directed graph, such that an edge between two nodes $e = (v,u) \in E$ means that user v mentioned user u during our collection period. We extract the 1.5 egocentric network $E_{1.5}(v)$, where v are the accounts in C_1 and C_2.

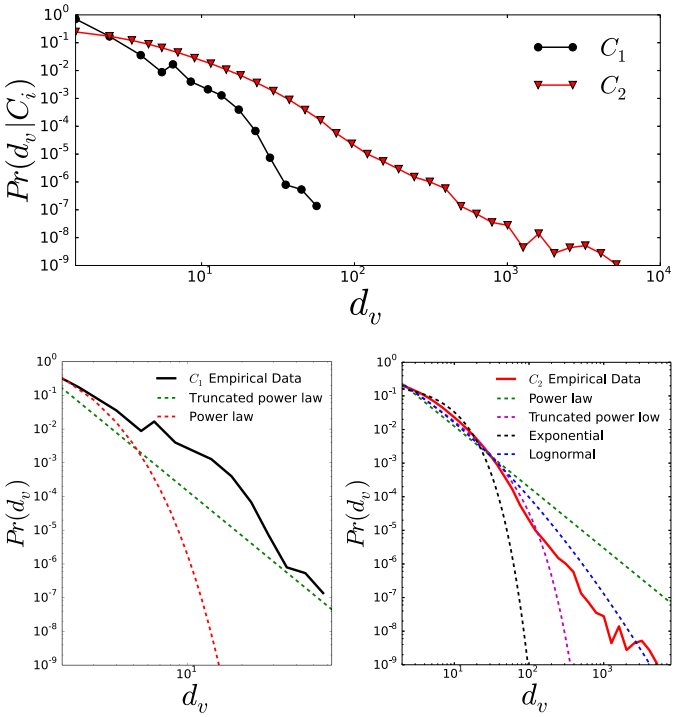


Figure 8: The top figure shows the distribution of the frequency of mentions $d(v)$ for C_1 (black circles) and C_2 (red triangles). The bottom figures compare the empirical distribution obtained with best fits of other heavy-tailed distributions (see Appendix A).

Figure 8 shows the degree distribution of the mention network. Although multiple studies observed that the degree for the mention network follows heavy-tailed distributions (e.g., [15]), in order to understand the topological structure, we further investigate the concrete goodness of fit (see Appendix A). The scale-free nature of the mention network (i.e., degree distribution that follows a power law) implies a very high heterogeneity level in user behavior, which is expected for human activity phenomena [4, 20]. In addition, the figure shows a clear difference between the length of the tail of the distributions between the two classes C_1 and C_2.

Table 3: Comparing different centrality measures for the mention network for C_1 and C_2 accounts

	Betweenness (C_B)		Closeness (C_C)	
Class	μ	σ	μ	σ
C_1	0.014	0.08	0.97	0.12
C_2	0.096	0.14	0.77	0.25

Table 3 compares two centrality measures for the mention network, namely the betweenness C_B and closeness C_C centralities. We observe that the average betweenness centrality for C_2 is significantly higher than C_1, which indicates that C_1 accounts target users that mention each other (i.e., communities and clusters of users). This is somewhat a surprising outcome, as we expect C_2 accounts to utilize the associated relationships of the original account owner, where the nodes in the neighborhood are real friends and are more likely to mention one another. The relatively low betweenness in C_1 can be explained by at least three possibilities:

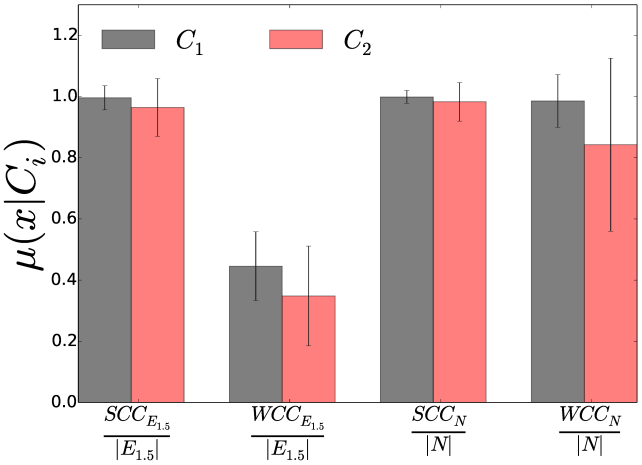


Figure 9: The density of connected components in the mention network for C_1 and C_2

- *Conversations hijacking*. We observe that 51.5% of the tweets captured by C_1 contain mentions, and 43.3% of these mentions are replies. In addition, only 1.2% of their mentions were reciprocated ($density_{bi} = 0.0127$), which arouses suspicion that C_1 accounts intrude on on-going conversations between legitimate users, and thus have resulted in a low betweenness centrality.

- *Targeting hubs*. Due to the scale-free nature (i.e., degree distribution that follows a power law) of the mention network, mentioning or replying to hubs (nodes that are highly connected to other nodes in the network) increase the chance that the alters will be connected, and hence the low betweenness score.

- *Crawling profiles*. It is also possible that C_1 accounts target communities and connected users in the mention graph by crawling profiles (i.e., visiting the followers/following lists or users' *timeline* of the seed targeted profile).

Figure 9 shows high average densities of strongly connected components for both the egocentric network and the neighborhood network in classes C_1 and C_2 (i.e., $\frac{SCC_N}{|N|}$ and $\frac{SCC_{E_{1.5}}}{|E_{1.5|}}$). This observation indicates a difficulty in forming reciprocal mention relationships as discussed earlier. Also, a higher score in the densities of weakly connected components ($\frac{WCC_N}{|N|}$ and $\frac{WCC_{E_{1.5}}}{|E_{1.5|}}$) for C_1 explains the lower betweenness centrality score observed in Table 3.

The discrepancy in network measures (i.e., degree distribution, centralities, and connectivity) between C_1 and C_2 indicates the existence of different strategies for reaching audiences employed by each class accounts.

7. RELATED WORK

We discuss prior related work on OSNs' spam and network analysis using the following categories: i) OSN organized spam campaigns; ii) OSN spam accounts analysis; and iii) spam detection in OSNs. Although we focus on spam accounts analysis, our first in its kind approach of spam behavioral categorization (i.e., identifying sub-populations), analyzing the different classes of spam accounts, and analyzing the mention interactions, all provide a unique view in looking at spam trends in OSNs.

7.1 Spam in Social Networks

With the rapid growth of OSNs popularity, we are witnessing an increased usage of these services to discuss issues of public interest and hence shape public opinions [8]. This model of users as an information contributors has provided researchers, news organizations, and governments with a tool to measure (to some degree) representative samples of populations in real time [1, 13, 17, 25]. However, Lumezanue et al. [16] identified *propagandists* Twitter accounts that exhibit opinions or ideologies to either sway public opinion, disseminate false information, or disrupt the conversations of legitimate users. The study focused on accounts connected to two political events: i) the 2010 Nevada senate race; and ii) the 2011 debt-ceiling debate. A similar campaign has been analyzed [26], in which spam accounts flood out political messages following the announcement of Russia's parliamentary election results. In addition, classical forms of abuse such as spam and criminal monetization exist in Twitter including phishing scams [6], spreading malware [22], and redirecting victims to reputable websites via affiliate links [27] to generate income.

7.2 Social Network Spam Analysis

Due to the popularity of social media services, several studies measured and analyzed spam in OSNs. Yang et al. [36] provided an analysis of some of the evasive techniques utilized by spammers, and discussed several detection features. In addition, Yang et al. [37] performed an empirical analysis of the social relationship in Twitter (i.e., following relationship) in the spam community. The study showed that spam accounts follow each other and form small-world networks. Stringhini et al. [23] examined *Twitter account markets*, and investigated their association to abusive behaviors and compromised profiles. Thomas et al. [28] performed a study in collaboration with Twitter to investigate the *fraudulent accounts* marketplace. The study discussed prices, availability, and fraud perpetrated by 27 merchants generating 127 to 459K US dollars for their efforts over the course of ten months. In another study [27], Thomas et al. examined tools, techniques, and support infrastructure spam accounts rely upon to sustain their campaigns. Surprisingly, the study showed that three of the largest spam campaigns in Twitter direct users to legitimate products appearing on reputable websites via affiliate links that generate income on a purchase (e.g., *Amazon.com*). However, the authors considered only tweets that contained URLs, and thus overlook malicious accounts that employ other spamming strategies, such as: i) embedding *non-hyperlink URL* by encoding the ASCII code for the dot; ii) *follow spam accounts* that generate large amounts of relationships for the hope the victim account would reciprocate the relationship or at least view the criminal's account profile which often has a URL embedded in its bio. Ghosh et al. [11] investigated the spammers' mechanism of forming social relationship (link framing) in Twitter, and found that vast majority of spam accounts are followed by legitimate users who reciprocate relationships automatically (social capitalists). The dataset used in this study contained 41,352 suspended Twitter accounts that posted a blacklisted URL. However, Grier et al. [14] discussed the ineffectiveness of blacklisting at detecting social network spam in a timely fashion.

7.3 Social Network Spam Detection

A number of detection and combating techniques proposed in the literature rely on machine learning. Benevenuto et al. [3] manually labeled 8,207 Twitter accounts, and developed a classifier to detect spammers based on the URL and hashtag densities, followers to following ratio, account-age, and other profile-based features. The account-age and number of URLs sent were the most discriminating features. Stringhini et al. [24] created a diverse set of "honey-profiles", and monitored activities across three different social networks (Facebook, Twitter, and MySpace) for approximately one year. They also built a tool to detect spammers on Twitter and successfully detected and deleted 15,857 spam accounts in collaboration with Twitter.

Another approach is presented by Xie et al. [35], where they designed and implemented a system that recognizes legitimate users early in OSNs. They utilized an implicit vouching process, where legitimate users help in identifying other legitimate users. Finally, Wanga et al. [34] investigated the feasibility of utilizing crowdsourcing as the enabling methodology for the detection of *fraudulent accounts*. This study analyzed the detection accuracy by both "experts" and "turkers" (i.e., workers from Amazon Mechanical Turk under a variety of conditions.

8. SUMMARY AND FUTURE WORK

This paper presents a unique look at spam accounts in OSNs through the lens of the behavioral characteristics, and spammers' techniques for reaching victims. We find that there exist two main classes of spam accounts that exhibit different spamming patterns and employ distinct strategies for spreading spam content and reaching victims. We found that C_2 (i.e. category 2 of spammers) and C_{normal} (i.e. legitimate users) manifest similar patterns across multiple attributes. We attempt to explain this observation through the hypothesis that C_2 mainly consists of compromised accounts, while the accounts in C_1 (i.e. category 1 of spammers) are fraudulent accounts, as we find support for the hypothesis throughout our analysis of profile properties. In terms of the relationship graph, we find that spam accounts that belong to C_1 are heavily skewed towards following rather than followers, which indicates difficulty in forming reciprocal relationships. Furthermore, we observe a low in-degree density for C_1, while C_2 has a more balanced in/out degree densities. We show that the betweenness centrality for C_1 in the mention graph is significantly lower than C_2, which might be a result of hijacking conversations, targeting hubs, or crawling profiles.

We acknowledge that our analysis may contain some bias. We have a partial view of the activities occurring during the data collection period due to the at most 1% sampling limit imposed by

Twitter. However, the work of Morstatter et al. [19] showed that the implications of using the Twitter Streaming API depend on the coverage and type of analysis. Generally, the streaming API can be sufficient to provide representative samples, that gets better with higher coverage, for certain types of analysis (i.e., top hashtags, topics, retweet network measures). Furthermore, we lack the absolute ground truth labels for the accounts presented in the dataset and primarily rely on Twitter's suspension algorithm. This might impose a lower bound on the number of spam accounts in our dataset (i.e., uncaught spam accounts are treated as legitimate users). In addition, there might be a fraction of legitimate users who deactivated their accounts during the collection period, and hence would be labeled as removed. We also lack the appropriate resolution for important attributes used in the analysis; for example, we only have the number of followers and following for each user, and not the actual relationships list. We also acknowledge that some of the explanations proposed in this work might lack rigorous validations, due to the difficulties in thoroughly obtaining the motivations and social actions of spam accounts. However, we believe that our first in its kind analysis of twitter functions and spam behavioral categorization describe well the current trends and phenomenon of OSN's spam and can be leveraged in designing OSN spam detectors and resilient architectures.

In our future work, we will design and test alternative labeling and validation mechanisms for the analyzed accounts. In addition, we plan to further investigate the differences between the spam accounts utilizing other interactions functions (e.g., hashtag, retweet, and favorite). We also intend to quantify the success of spam campaigns and explore the tools, techniques, and spam underground markets utilized by spam accounts to spread their content and evade many of the known detection mechanisms.

Acknowledgments

The authors would like to thank King Abdulaziz City for Science and Technology (KACST) for funding this work. In addition, the authors thank the Center for Complex Engineering Systems (CCES) at KACST and MIT for their support.

9. ADDITIONAL AUTHORS

Additional authors: Abdulrahman Alarifi (King Abdulaziz City for Science and Technology, email: aarifi@kacst.edu.sa), Anas Alfaris (Center for Complex Engineering Systems, email: anas@mit.edu), and Alex (Sandy) Pentland (MIT - Media Lab, email: pentland@mit.edu).

10. REFERENCES

[1] A. Almaatouq, F. Alhasoun, R. Campari, and A. Alfaris. The influence of social norms on synchronous versus asynchronous communication technologies. In *Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia*, PDM '13, pages 39–42, New York, NY, USA, 2013. ACM.

[2] Y. Altshuler, N. Aharony, A. Pentland, Y. Elovici, and M. Cebrian. Stealing reality: When criminals become data scientists (or vice versa). *IEEE Intelligent Systems*, 26(6):22–30, 2011.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on Twitter. In *Proceedings of the Seventh Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, July 2010.

[4] J. Borondo, A. J. Morales, J. C. Losada, and R. M. Benito. Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish Presidential election as a case study. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(2), 2012.

[5] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177, 2001.

[6] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru. Phi.sh/$ocial: The phishing landscape through short urls. In *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, CEAS '11, pages 92–101, New York, NY, USA, 2011. ACM.

[7] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, Nov. 2009.

[8] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. Political polarization on twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[9] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna. COMPA: Detecting Compromised Accounts on Social Networks. In *ISOC Network and Distributed System Security Symposium (NDSS)*, 2013.

[10] L. C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41, Mar. 1977.

[11] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 61–70, New York, NY, USA, 2012. ACM.

[12] GlobalWebIndex. Global web index: Q4 2012, 2013.

[13] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, and Y. Moreno. The dynamics of protest recruitment through an online network.

[14] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The underground on 140 characters or less. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, pages 27–37, New York, NY, USA, 2010. ACM.

[15] S. Kato, A. Koide, T. Fushimi, K. Saito, and H. Motoda. Network analysis of three twitter functions: Favorite, follow and mention. In D. Richards and B. Kang, editors, *Knowledge Management and Acquisition for Intelligent Systems*, volume 7457 of *Lecture Notes in Computer Science*, pages 298–312. Springer Berlin Heidelberg, 2012.

[16] C. Lumezanu, N. Feamster, and H. Klein. bias: Measuring the tweeting behavior of propagandists. In *ICWSM*, 2012.

[17] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 227–236, New York, NY, USA, 2011. ACM.

[18] M. McCord and M. Chuah. Spam detection on twitter using traditional classifiers. In *Proceedings of the 8th international conference on Autonomic and trusted computing*, ATC'11, pages 175–186, Berlin, Heidelberg, 2011. Springer-Verlag.

[19] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the sample good enough? comparing data from Twitter's streaming API with Twitter's Firehose. *Proceedings of ICWSM*, 2013.

[20] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, December 2005.

[21] H. Nguyen. 2013 state of social media spam. Technical report, Nexgate, 2013.

[22] A. Sanzgiri, A. Hughes, and S. Upadhyaya. Analysis of malware propagation in twitter. *Reliable Distributed Systems, IEEE Symposium on*, 0:195–204, 2013.

[23] G. Stringhini, M. Egele, C. Kruegel, and G. Vigna. Poultry markets: On the underground economy of twitter followers. In *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks*, WOSN '12, pages 1–6, New York, NY, USA, 2012. ACM.

[24] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. *Proceedings of the 26th Annual Computer Security Applications Conference on - ACSAC '10*, page 1, 2010.

[25] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in twitter events. *J. Am. Soc. Inf. Sci. Technol.*, 62(2):406–418, Feb. 2011.

[26] K. Thomas, C. Grier, and V. Paxson. Adapting Social Spam Infrastructure for Political Censorship. In *Proceedings of the 5th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, Apr. 2012.

[27] 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*, IMC '11, pages 243–258, New York, NY, USA, 2011. ACM.

[28] 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 Security Symposium*, 2013.

[29] Twitter. Following rules and best practices. `https://support.twitter.com/articles/68916-following-rules-and-best-practices`, 2012. [Online; accessed 22-October-2013].

[30] Twitter. Public stream. `https://dev.twitter.com/docs/streaming-apis/`, 2012. [Online; accessed 1-October-2013].

[31] Twitter. Rules. `https://support.twitter.com/articles/18311-the-twitter-rules`, 2012. [Online; accessed 1-October-2013].

[32] Twitter. Initial public offering of shares of common stock of twitter, inc. `http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm`, 2013. [Online; accessed 5-October-2013].

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

[34] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. J. Metzger, H. Zheng, and B. Y. Zhao. Social turing tests: Crowdsourcing sybil detection. In *NDSS*. The Internet Society, 2013.

[35] Y. Xie, F. Yu, Q. Ke, M. Abadi, E. Gillum, K. Vitaldevaria, J. Walter, J. Huang, and Z. M. Mao. Innocent by association: Early recognition of legitimate users. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 353–364, New York, NY, USA, 2012. ACM.

[36] C. Yang, R. Harkreader, and G. Gu. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In R. Sommer, D. Balzarotti, and G. Maier, editors, *Recent Advances in Intrusion Detection*, volume 6961 of *Lecture Notes in Computer Science*, pages 318–337. Springer Berlin Heidelberg, 2011.

[37] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 71–80, New York, NY, USA, 2012. ACM.

[38] C. M. Zhang and V. Paxson. Detecting and analyzing automated activity on twitter. In *Proceedings of the 12th international conference on Passive and active measurement*, PAM'11, pages 102–111, Berlin, Heidelberg, 2011. Springer-Verlag.

APPENDIX

A. THE SCALE-FREE NATURE OF THE MENTION NETWORK

We investigate the scale free nature of the mention network by examining whether power law is the best description for our data's degree distribution. We achieved this by comparing the power law fit to fits of other distributions using log-likelihood ratios R and generating p-value p (the significance for this ratio) to specify which fit is better [7] (see Table 4 and Figure 8). Generally, the first distribution is a better fit when $R > 0$, alternatively the second distribution should be preferred when $R < 0$. We find for C_2 is significantly (with $p = 1.8^{-173}$) best described as a truncated power law distribution. As for the case of C_1 power law is insignificantly better describer than truncated power ($p = 0.9$).

Table 4: Comparing different heavy-tailed distributions for the degree distribution of the mention network.

Candidates	**Class**	R	p
Power law vs Exponential	C_1	193.8	$< 10^{-10}$
Power law vs Trunc. power law	C_1	0.03	0.9
Power law vs Exponential	C_2	45.7	$< 10^{-10}$
Power law vs Trunc. power law	C_2	-68.2	$< 10^{-10}$
Power law vs Lognormal	C_2	-111	$< 10^{-10}$
Trunc. power law vs Lognormal	C_2	28.1	1.8^{-173}