DETECTING SUBVERSION ON TWITTER

VLADISLAV KONTSEVOI, NAIM LUJAN, AND ADRIAN OROZCO

Abstract. Twitter is a dominant platform enabling millions of users to self-organize and communicate. Its popularity has attracted spammers seeking to capitalize on its infrastructure and user base. Indeed, Twitter has useful features that enable users to discover and comment on trending topics. Spammers can undermine the integrity of these features and exploit high-traffic topics, detracting from otherwise valuable discussions. In general, forged Twitter accounts negatively impact the user experience and must be dealt with.

In this paper, we examine existing methods for detecting unwanted activity on Twitter and present a state-of-the-art system for identifying Twitter spammers. We gathered an appropriate dataset using Twitter's APIs. Synthesizing ideas from the literature, we proposed options to improve our system's performance. We specified certain metrics ("features"), calculated on gathered data, that assist in the classification of user accounts. The metrics fall into three primary categories: profile-based features, Tweet-based features, and network-based features. Despite having insufficient resources and data to draw strong conclusions at the time of writing, we expect to finish our research in the near future.

1. Introduction

Twitter is one of the most popular social networking platforms in the world, with millions of active users. Its users routinely reveal sensitive information about their lives. They implicitly rely on Twitter's security to feel comfortable sharing such information. They also rely on the platform to consume new information. Twitter's value rests on the integrity of its users and of every contributor to the social platform. If some accounts are illegitimate, then users' information can be compromised and exploited. Adversaries can send spam to users, try to direct them to malicious websites and scams, and unduly influence them.

Adversaries can also hack legitimate accounts and use them for malicious purposes. These types of attacks are harder to detect, because hacked accounts are originally legitimate, and their characteristics are not so different than those of legitimate accounts.

The growth of social media has given scammers the ability to reach wide audiences with ease. A scammer can post to Twitter and wait for users to inadvertently visit their scam. A certain percentage of visitors may even be victims of the scam. Popular events that gain national attention are

Date: May 14, 2014.

especially targeted by scammers who realize that trending topics will receive substantial user traffic. A prime example occurred in the aftermath of the 2013 Boston Marathon Bombing: a malicious Twitter account was created shortly after the initial news report and promised to donate $1 for every retweet. However, the supposed charity was a scam that preyed on the kindness of the public [2]. Such a breach in Twitter's integrity causes users to lose confidence in the Twitter social network and is detrimental to all honest users.

In addition, many companies base their business models on the marketability of user information. Fake accounts can falsely increase users' popularity, compromising business models that use user popularity. Thus, it is very important for social media websites to automate the detection and removal of fake accounts. Unfortunately, this problem has not as yet been fully solved.

In this paper, we investigate a technique for detecting subversion on Twitter. We programmatically crawl the Twitter social graph. We define a set of features that may be computed on our data which can help predict whether or not a user is a spammer. Finally, we apply supervised learning techniques to enable us to determine whether a user is a spammer. Our end result is a research system that can sample the Twitter social graph, compute interesting features on the data, and classify users as being spammers or not using machine learning techniques.

2. Background and Related Work

Social networks provide users with a platform for connecting and interacting with others. This is a valuable system for people, businesses, and groups to organize their relationships and communicate. For example, the Red Sox can use Twitter to broadcast inning by inning highlights, final scores, and significant news about players or the league. Businesses can use Twitter to quickly announce promotions or news to their customer base instantaneously and free of charge. Given Twitter's ability to impact so many, users and organizations are always eager to reach wider audiences. This incentive leads them to take extreme measure to reach more people, often resorting to subversive methods to expand their following.

2.1. **Twitter.**

Twitter is an information sharing system designed as a microblogging platform where users send short text messages(Tweets) that appear on their friends' pages. Users can follow others and receive information along their social links. Unlike other social media websites such as Facebook and Myspace, relationships are directional: a user has both followers and followings, as opposed to bidirectional links that represent true friendship. A user only sees the Tweets of the users she is following; likewise, her followers are the only users who can see her Tweets.

Unlike other social networking websites, no personal data is shown on a Twitter profile by default. By default, a user will show a username and

optionally a real name as well. Users' profiles are initially configured to be public, although users can decide to protect their profiles, requiring acquaintances to ask for explicit permission before they can start following.

Tweets can be repeated by a user through a "Retweet." A Retweet shows the Tweet of another user along with a "RT @username" at the beginning, citing the user who posted the original Tweet. Retweet spam can exist where legitimate links are changed to illegitimate ones that are obfuscated by URL shorteners.

Users use hashtags (#) to identify topics relevant to Tweets. For example, using the hashtag "#worldcup" indicates that a Tweet is related to the World Cup. Any user can query Tweets related to a given topic by searching for a specific hashtag. The most popular hashtags are displayed in the "trending topics" page. Such topics usually are shocking and involve breaking news such as the Boston Marathon Bombing, Iranian elections, Sandy Hook shooting, etc. The most trending hashtags can be abused by spammers who post completely unrelated Tweets in association with the hashtag [6].

2.2. **Related Work.** Among the biggest security concerns in social media is the creation of faked and cloned accounts. In creating such accounts, attackers gain the trust of others, thereby luring victims into clicking links contained in messages leading to phishing or drive-by downloading websites. Many users are oblivious to the existence and abundance of malicious accounts and are easily exploited. Even today, social networking websites lack automated systems to detect fake accounts because it is very difficult to reliably capture the diverse behavior of fake and real online social network profiles.

Using hacked accounts to send spam can be viewed as an instance of a Sybil attack on a social network. This involves attackers controlling a large number of counterfeit accounts on a network, giving the attackers an unduly strong influence in the system. For example, a user with control over many artificial Reddit accounts can use them to upvote or downvote certain posts with more weight than a single user is meant to have on the website. Twitter is particularly vulnerable to these types of attacks given the streamlined and lenient process of account creation. As a website that centers around users following one another, Twitter's worth is also significantly impacted by false followers and false Retweets.

We examine two types of Sybil attacks: one involves creating many fake accounts, and the other employs hijacked legitimate accounts to perform a desired function. As mentioned, the first attack is particularly problematic for Twitter, although fake accounts are relatively easy to detect based on their behavior. Upon detection, the accounts may be removed to keep Twitter spam-free. An attacker may respond to such a defense by hacking into legitimate users' accounts and using them to carry out their attack. The

compromised accounts, being legitimate for the most part, are not readily identifiable, as their behavior is less divergent from that of a real user [9].

There are many approaches to detecting compromised or false user accounts. Some approaches attempt to systematically rank potentially dishonest accounts by their probability of being dishonest; others use supervised learning algorithms, while others still use social honeypots. One approach [10] to detecting cloned profiles on various social media platforms involves extracting information from a legitimate account and querying the Internet with information from that account. Based on the number of results, user information is categorized as common or user-specific. Account data are then queried to find possible profiles that are clones of the original. After obtaining a list of possible clones, each is examined and given a similarity score in relation to the genuine account. At the end of the process, the user is presented with all possible clones ranked by similarity score, which are then processed with human intervention.

Another approach to detecting fake accounts, called SybilRank [7], relies on properties of the social graph to rank users by their perceived likelihood of being fake. SybilRank uses the observation that an early-terminated random walk starting from a known legitimate user has a higher normalized probability of landing on a legitimate user than on an illegitimate user. Cao et al. also argue that human intervention is necessary to determine whether a user is a spammer to prevent unacceptable false positives.

A different approach to detecting social spammers involves social honeypots [11]. Honeypots are created to trap attackers and begin monitoring and logging attackers' activity. When a honeypot's profile receives an unexpected friend request, the user sending the request is put under observation. The user's activity is tracked for later use as evidence by a classifier to decide if the suspected user is a spammer or not. In this way, the authors identify spammers with a low false positive rate and are able to identify characteristics of spammers' profiles that can be applied to detecting previously unknown spammers.

The final class of approaches to detecting spammers assigns users a vector of values ("feature" values) [14, 6, 11, 13] capturing different attributes of their profiles, Tweet history, local social graph, etc. Features are carefully constructed using empirical user data, before they are used as input to a supervised machine learning algorithm along with a set of users that have been pre-classified as being spammers or not. The resulting classifier can then be used on the broader social network.

Approaches based on machine learning can be augmented using statistical analysis of the language used in Tweets [12]. For example, a suspicious Tweet about a trending topic can be compared to the broader thread of Tweets about the topic using the concept of Kullback–Leibler divergence. Augmenting typical profile-based features with sophisticated Tweet-based features can improve the performance of a spammer-classification system.

Our work uses a machine learning framework with profile- and Tweet-based features. We synthesize diverse ideas from the literature and describe our proof-of-concept system that samples the Twitter social graph, computes interesting features on the data, and classifies users as being spammers or not.

3. Data Collection

There are several approaches to collecting useful data from Twitter. Some authors asked Twitter to white-list their servers to enable the collection of larger quantities of data than would otherwise be possible [6]. Others rely on indirect tools of collecting data such as social honeypots [11]. Others still use the Twitter Streaming and REST APIs [14, 11]. Some authors combine multiple techniques for gathering data. In our data collection, we combine the Streaming and REST APIs. Our goal was to crawl approximately 100000 users, about $0.5\% - 1\%$ of which were expected to be spammers.

3.1. **Twitter API.** To gather data, we combined the Twitter Streaming API [5] and the Twitter REST API [4]. The REST API was the limiting factor, with surprisingly restrictive rate limits. In particular, the API allows an application to retrieve one list of up to 5,000 followers or followings per minute. With one API key, crawling 100,000 users and retrieving followers/followings would take almost 70 days. Even with 20 API keys, it would take 3.5 days of continuous queries, not accounting for network problems.

The REST API limit on retrieving recent Tweets, which is necessary to establish the ground truth of whether a user is a spammer, is less restrictive. The API allows an application to retrieve 12 lists of Tweets per minute. With one API key, crawling 100,000 users would take about 5.75 days. With 20 API keys, it would take about 7 hours. In practice, it took a full day of continuous monitoring.

We considered building a web-based crawler to circumvent the API, but we believed that this would be prohibitively time-consuming.

3.2. **Crawling and Ground Truth.** We roughly follow the approach used in [14]. We first use the Twitter Streaming API to randomly sample a set of recent Tweets, from which we determine 15 unique Twitter users. For these 15 seed users, we use the Twitter REST API to collect data including the screen name, real name, follower count, friend count, creation date, profile description, favorite Tweet count, and tweet count. We also sample each user's 200 most recent Tweets, as well as up to 5,000 users that follow or are being followed by the given user. We repeat this seeding process several times.

We then crawl the followers and followings of our seed users. The users crawled in this second round greatly outnumber the users crawled in the seed round. Due to unfortunate rate limitations in the Twitter REST API, we do not collect followers or followings for users crawled in the second round.

During this round, we discard users whose Tweets are not public. Such users are unlikely to be spammers, as spammers prefer exposure, although they may belong to a second class of more personal malicious actors.

After gathering a sufficient quantity of user data, we determine the ground truth for whether or not a user is a spammer. We again follow the approach of [14], using the Google Safe Browsing API [1] to determine whether links posted in Tweets are malicious. If a Tweet contains at least one malicious URL, we call it malicious. If at least one crawled Tweet is identified as being malicious, we manually examine the user's remaining Tweets to determine whether that user as being a spammer.

Unfortunately, using the Google Safe Browsing API is insufficient to identify all spammers. Spammers sometimes use `bit.ly` redirection URLs, which feature automatic malware and spam detection. When a malware or spam link is posted using a `bit.ly` redirection URL, it is automatically redirected to a warning page. It is possible to detect that a redirection to a warning page occurs, but it takes about 0.5 seconds to do so. Doing this for upwards of 2,500,000 URLs, an appreciable portion of which are from `bit.ly`, is infeasible. Instead, we examined a cross-section of about 25,000 users for malicious redirection URLs.

Our dataset may contain bias, in the form of false positives: users inadvertently posting several malicious Tweets without being spammers themselves. Our data may also contain false negatives: users who we label as being legitimate, yet are spammers who do not post many malicious URLs or post malicious `bit.ly` redirection URLs that we do not identify. Even with these potential biases, we can use the dataset to study the effectiveness of supervised learning techniques for catching spammers.

Summary statistics about our dataset are found in Table 1. We note that the users sampled are prolific tweeters and tend to use more hashtags than URLs. The sampled users correspond to the followers and following of the seed set of users. Because users are likely to follow accounts with many followers (having many followers is "sticky"), the sampled users have relatively many followers. We note that we were only able to identify 34 spammers using the Google Safe Browsing API and with our limited use of `bit.ly` redirection. We could have identified about 120 in total with more extensive use of redirection. This is still less than the 450 or so that we had expected to identify by extrapolating the results of [14].

4. Feature Engineering

We define three broad classes of features: profile-based features that use characteristics of a Twitter user's profile, Tweet-based features that use characteristics of Tweets, and network-based features that use characteristics of the Twitter social graph near a specific user. We do not use all features described in the present work; in particular, we do not use any network-based features.

Statistic	Value
Users in dataset	110,860
Identified spammers	34
Total following	642,303,936
Total followers	5,718,689,957
Total Tweets	1,089,383,677
Tweets in dataset	17,865,631
Hashtags in dataset	4,593,333
URLs in dataset	3,570,342

TABLE 1. Dataset summary statistics.

4.1. **Profile-based Features.** Profile-based features, sometimes called user behavior attributes [6], capture characteristics of a user's profile that may be predictive of whether or not the user is legitimate. In our modeling, we include simple features such as the number of followers and following, the number of favorited Tweets, and so on. We also include the more sophisticated features described in this section. For instance, we include features that capture the effects of automation, as recommended in [14].

Following/Follower Ratio. On average, spammers have more followers than they have users following them, because it is much easier to follow a user than it is to gain a follower. Of course, there are exceptions to this: for instance, spammers may reciprocally follow one another and form clusters in the social network. We introduce two features to measure this hypothesized effect, the followings/followers balance

$$\text{balance} = \frac{\#(\text{followings}) + 1}{\#(\text{followings}) + \#(\text{followers}) + 2},$$

and the log-ratio of followings to followers:

$$\text{logRat} = \log\left(\frac{\#(\text{followings}) + 1}{\#(\text{followers}) + 1}\right).$$

Both of these features are intended to capture this phenomenon in slightly different ways. For example, the log-ratio of followings to followers will be quite small for a popular legitimate account, but quite large for an illegitimate account that indiscriminately follows others.

Account Age. As compared to newer accounts, older accounts are less likely to belong to spammers, as spammers are likely to be caught and to quickly create new accounts. Thus, we use the age of the account as a feature in detecting spammers.

Tweeting Rate. The number of Tweets tweeted per day by a spammer may be different than that of a legitimate user. To investigate the extent to which is true, we use the tweeting rate as a feature in our system.

Following Rate. Spammers are likely to quickly follow users after creating their accounts than legitimate users are. We use the following rate (in users followed per day) as a feature to capture this phenomenon.

Follower Rate. An average spammer is likely to have fewer followers per day elapsed after account creation than an average legitimate user does. To capture this, we define the follower rate in followers per day.

API Ratio. Because spammers often choose to use the Twitter API to post tweets [8], we consider the proportion of Tweets that are posted from the API. We hypothesize that a high API ratio is predictive of a user being a spammer. It may also indicate that the user is using a non-native client to post Tweets.

Tweeted to Favorited. We hypothesize that legitimate users are likely to "favorite" others' Tweets, while spammers are likely to post Tweets without consuming content. Thus, we examine the normalized ratio of own Tweets tweeted to others' Tweets favorited:

$$\text{favToTweet} = \log\left(\frac{\#(\text{tweeted})+1}{\#(\text{favorited})+1}\right).$$

4.2. **Tweet-based Features.** Tweet-based features, also called content attributes [6], capture properties related to the way that users write Tweets: that is, the way that users communicate with one another. In examining Tweets, we can consider quantitative characteristics including Tweet length, special characters used, and metrics on individual Tweets. We can also consider qualitative markers that can be extracted and summarized at the user level to help inform account classification. For example, a common mode of spam involves embedding URLs into Tweets that link to websites or advertisements unrelated to the rest of a Tweet's content. This kind of attack requires a URL to be used in the message, so measuring the presence of URLs may be helpful for detecting spammers.

URL Ratios. Spammers are more likely to tweet URLs as compared to non-spammers [6]. Thus, we define two features to capture this phenomenon: average URLs per Tweet and average URLs per word tweeted. We expect that both of these metrics will be higher in spammers than in non-spammers.

Hashtag Ratios. We expect that spammers and non-spammers communicate differently using hashtags. Specifically, spammers often use hashtags to appear in the public feed for trending topics. We define two features to study the usage of hashtags by Twitter users: average hashtags per Tweet, and average hashtags per word tweeted.

Average Tweet Length. Legitimate users share their thoughts and details about their lives, while illegitimate spammers share spam links. We hypothesize that legitimate users post longer Tweets on average.

Spam Ratios. Spammers are far more likely to use "spam" words in their Tweets. Thus, it may be interesting to study the proportion of Tweets that contain certain spam words and phrases.

Username Mentions. Spammers are unlikely to mention others' usernames (e.g. @Vlad2014) in their Tweets [11]. As a proxy for this, we define a feature that captures the number of at signs (@) per Tweet.

API URL Ratio. As mentioned, spammers often use the Twitter API to post tweets, including Tweets with malicious URLs and other spam. We measure the extent to which users use the API to post URLs by considering the ratio of the number of URLs posted using the Twitter API to total URLs posted.

Favorites and Retweets per Tweet. A user who produces high-quality content is not likely to be a spammer. Users who recognize the value of a Tweet will either add it to their list of favorites, or retweet (rebroadcast) it. Thus, if a user has a high proportion of favorites or retweets, he or she is not likely to be a spammer.

Proportion of Retweeted Tweets. Users who retweet others' Tweets produce less original content and are more likely to be spammers. We capture this by measuring the proportion of Tweets that are Retweets of others' content. Retweets are typically identified by the prefix "RT @".

Tweet Similarity. Spammers tend to tweet the same or very similar messages over a span of time. That is, unlike a legitimate user posting about a broad range of real-life occurrences, spammers often post the same advertisements or malicious URLs. To measure this phenomenon, we can quantify the similarity of users' Tweets to one another.

4.3. **Network-based Features.** A powerful third class of features comes from the theory of networks. Although spammers can choose who to follow and may even attract followers of their own, they cannot in general influence relationships between their neighbors. For this reason, network-based features can be very powerful. Despite their power, applying them is infeasible due to Twitter API limitations. We include descriptions and motivations for the more interesting network-based features, many of which are based on [14]. Mathematical formulas are omitted.

Local Clustering Coefficient. On a social network, the local clustering coefficient measures the degree to which a node's neighbors are themselves interconnected. A high local clustering coefficient indicates that a node is part of a tightly-knit group. In the case of Twitter, if a large proportion of user's followings follow one another, then the user is likely to be legitimate.

Bi-directional Links Ratio. When two users follow one another, they are said to have a bi-directional link between one another. Personal acquaintances are likely to have bi-directional links to a user, and a high proportion of bi-directional links indicates that many of a user's followers are likely to be personal acquaintances. Spammers are likely to have a low number of bi-directional links, as a fraction of total followings.

PageRank. The PageRank is a measure of the influence of a user and accounts for indirect contributions to influence. That is, if a user is followed by many users who themselves have many followers, the user will have a relatively higher PageRank. The PageRank may be approximately computed on a sample of a social network, but computing it on the Twitter social graph is difficult because we are only able to crawl followers and followings for a small seed set of users.

Average Neighbors' Followers. On average, legitimate users follow accounts who themselves have more followers because such accounts are of higher quality. Thus, the average number of followers of a user's followings measures the quality of following choices made by a user. This quantity is expected to be lower in spammers who follow indiscriminately. However, this quantity can be gamed by an adversary who is aware that we are using this metric.

5. Evaluation

We first present statistical and graphical information about the data gathered. We then present our present work towards applying our dataset to detect spammers.

5.1. **Data Exploration.** We present means of certain features for presumed legitimate Twitter users and for spammers in Table 2. First, we notice that legitimate accounts tend to have more followers than do illegitimate accounts as hypothesized. Meanwhile, we cannot draw any useful conclusions about the number of followings of legitimate and illegitimate accounts. Because the distributions of followers and followings are right-skewed, it may be interesting to consider them on a logarithmic scale.

Next, we note that legitimate accounts tend to favorite far more Tweets than do illegitimate accounts. Of course, account age is a confounding factor, but our experiments reveal that this difference is still statistically significant when normalizing by account age, motivating a potential new feature.

We notice that the following-follower log-ratio is greater in the case of spammers as expected. This is further shown in Figure 1. In addition, the Tweeting rate is significantly higher for legitimate users. Legitimate users use Twitter throughout the day, while illegitimate users post periodic messages of little value. The empirical CDF for this feature can be seen in Figure 2. We see that the most prolific legitimate users tweet far more often than spammers do.

Because spammers can choose when to follow others, they can spoof the following rate. We see that the following rates are close for legitimate users and for spammers. Likewise, and somewhat disappointingly, we see that spammers can spoof the tweeted to favorited ratio. Meanwhile, it is harder for them to spoof the following rate, which may be used to identify spammers.

We note that spammers use more URLs per Tweet than do legitimate users, as expected. The cumulative density functions of this feature are shown in Figure 3. However, differences in hashtag usage are barely discernible. Differences in the frequency of Retweets are not discernible at all. On the other hand, Tweets made by legitimate users are favorited more often by others, as shown in Figure 4.

Feature	Mean (Legitimate Users)	Mean (Spammers)
NumFollowers	51598.34 ± 3704.97	7420.82 ± 10225.51
NumFollowing	5793.10 ± 202.42	8182.79 ± 10900.86
FavoritesCount	1228.41 ± 35.92	69.59 ± 50.72
FllingFllwerLogRatio	-0.03 ± 0.01	1.07 ± 0.53
TweetingRate	13.69 ± 0.19	2.56 ± 1.21
FollowingRate	9.81 ± 0.26	8.21 ± 7.49
FollowerRate	43.74 ± 2.25	6.17 ± 6.94
TweetedToFavorited	3.13 ± 0.01	4.23 ± 0.87
URLPerTweet	0.19 ± 0.00	0.57 ± 0.14
HashtagsPerTweet	0.26 ± 0.00	0.42 ± 0.19
FavoritesPerTweet	27.76 ± 4.42	0.16 ± 0.17
RetweetRatio	0.27 ± 0.00	0.28 ± 0.10

TABLE 2. Feature means, with 95% confidence intervals.

5.2. **Detecting Spammers.** Unfortunately, we did not have a sufficient number of spammers in our dataset to be able to train a robust classifier. We did train an SVM classifier using a radial basis kernel that performed adequately on a small holdout set.

6. Limitations and Future Work

We only crawl a subset of the Twitter social graph, and this subset may have sampling bias. Our subset does not include links present between users because of Twitter REST API limitations. Unfortunately, we cannot access better data without having a higher level of privilege that we were unable to obtain.

Furthermore, it is difficult to determine the ground truth for Twitter spammers without human intervention. Human resources were limited for the current research deadline, and we were not able to implement sophisticated techniques for determining whether a user is a spammer, such as using

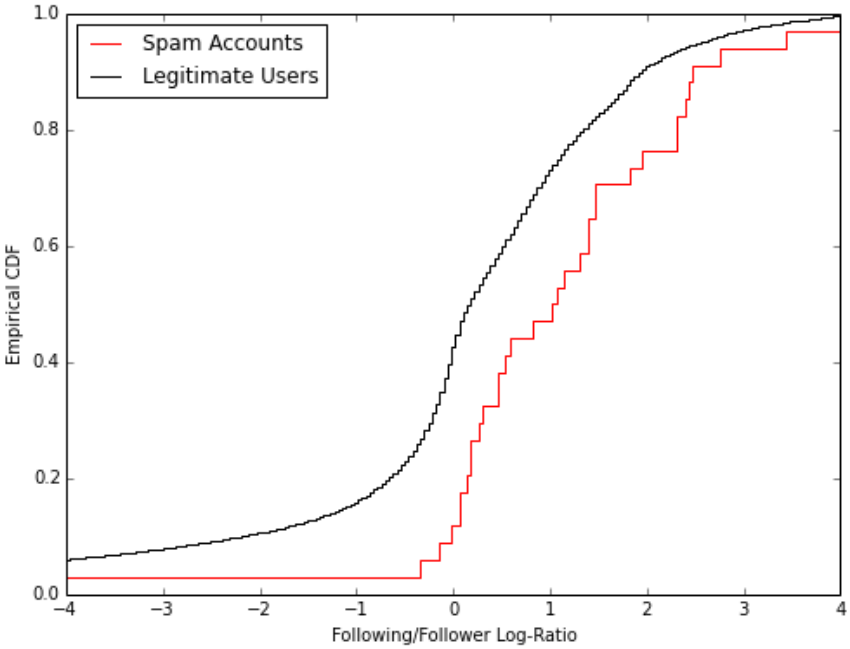

FIGURE 1. Empirical CDF of following/follower log-ratio.

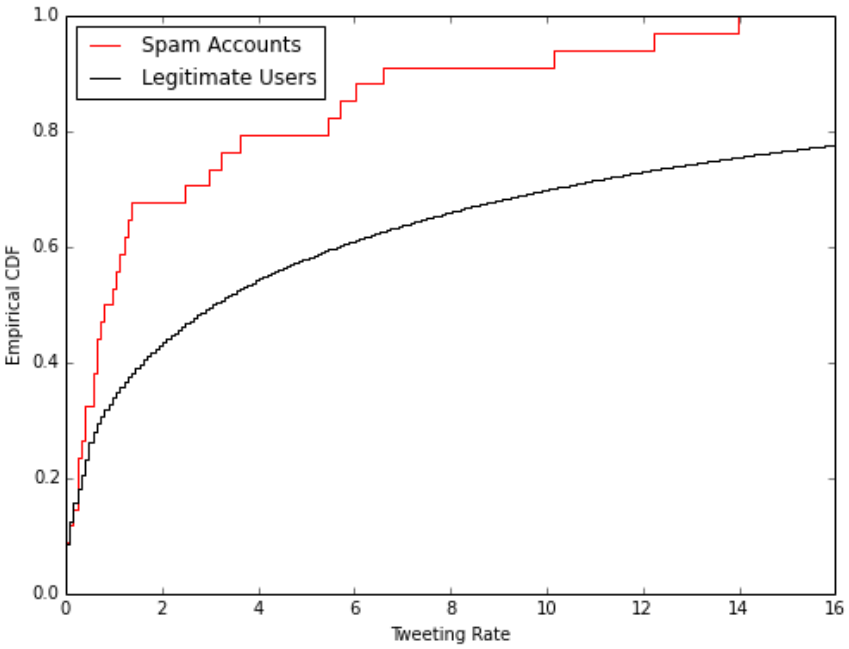

FIGURE 2. Empirical CDF of tweeting rate.

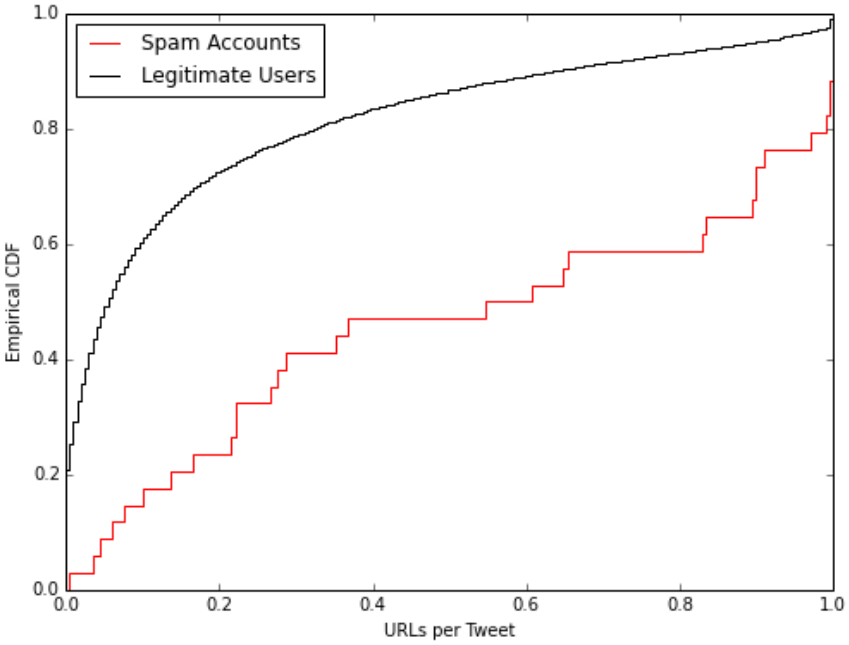

FIGURE 3. Empirical CDF of URLs per Tweet.

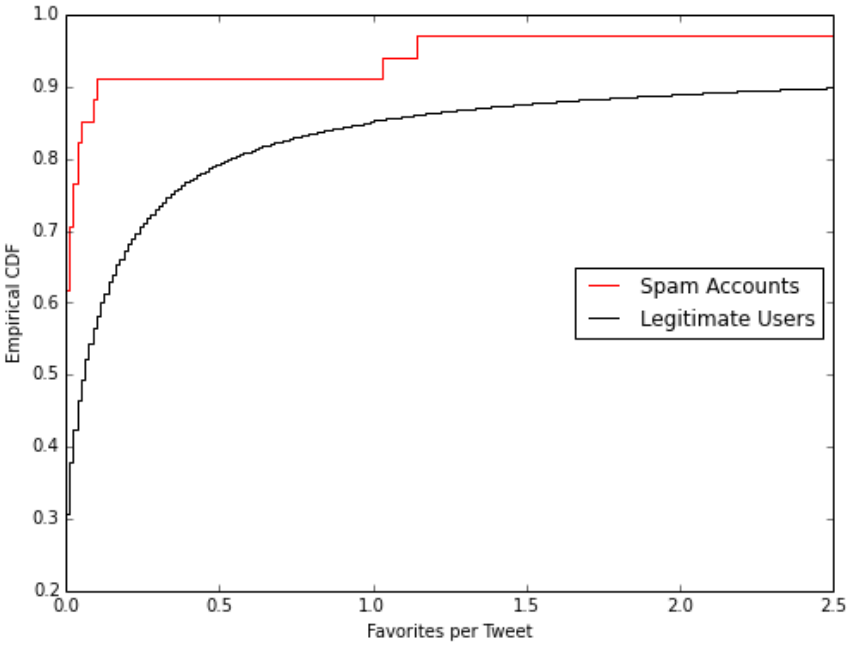

FIGURE 4. Empirical CDF of favorites per Tweet.

Capture-HPC [14] or performing full `bit.ly` redirection. Nevertheless, we believe that our ground truth has an acceptably low false positive rate and allows us to draw useful conclusions. Unfortunately, we sampled a small number of spammers relative to the total number of users sampled.

Due to limitations in our dataset and in the Twitter REST API, we were unable to calculate many interesting graph-based features. We believe that these features are especially useful in determining whether a user is a spammer. However, we were able to use appropriate profile-based and Tweet-based features.

In potential future work, we would like to work directly with Twitter to obtain a better level of access to their dataset. We would also like to spend more time determining the ground truth for Twitter spammers. Additionally, we would like to apply feature selection techniques to determine which features are significant and to apply more supervised classification techniques. Finally, we would like to apply unsupervised learning techniques (e.g. clustering) to our dataset to see if we observe any interesting patterns.

We would also like to work with other datasets that may be more amenable to academic research. One interesting dataset is that of Quora [3], a knowledge sharing community. Like on Twitter, users have asymmetric followers and followings. While many users on Quora post interesting questions and answers, others post low-quality content. It would be useful to have an automatic means of detecting and de-prioritizing such users based on their profile, posting, and voting characteristics.

7. Conclusions

Twitter's simple design and features make it especially vulnerable to attackers because it is relatively easy for users to make accounts and post messages that can be viewed by many others. The problems caused by illegitimate account holders on social media sites like Twitter have not yet been solved. In our research, we studied current approaches to combating spam on Twitter.

We attempted to solve the problem of detecting illegitimate users on Twitter using a supervised learning approach: defining features based on an intuitive understanding of the behavior of legitimate and illegitimate users and using these features as inputs to a supervised classification algorithm. Although we gathered a large number of user accounts, we were not able to identify sufficiently many confirmed spammers from among these accounts.

We intend to continue this project, gathering additional data and refining our understanding of Twitter spammers to engineer better features. We hope that we will be able to apply network-based features as well, which we expect to be quite powerful.

Acknowledgements. We are grateful to Ronald Rivest and the other 6.857 course staff for giving us advice and for giving us the opportunity to work on this project.

References

1. *Google safe browsing API*, `https://developers.google.com/safe-browsing/`, 2013.
2. *In God we trust*, `http://waverleycab.org.uk/05/campaigns/in-god-we-trust/`, 2013.
3. *Quora*, `http://quora.com/`, 2014.
4. *Twitter rest API*, `https://dev.twitter.com/docs/api/1.1`, 2014.
5. *Twitter streaming API*, `https://dev.twitter.com/docs/api/streaming`, 2014.
6. F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, *Detecting spammers on twitter*, Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS), July 2010.
7. Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro, *Aiding the detection of fake accounts in large scale social online services*, Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (Berkeley, CA, USA), NSDI'12, USENIX Association, 2012, pp. 15–15.
8. Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia, *Who is tweeting on twitter: Human, bot, or cyborg?*, Proceedings of the 26th Annual Computer Security Applications Conference (New York, NY, USA), ACSAC '10, ACM, 2010, pp. 21–30.
9. Manuel Egele, Gianluca Stringhini, Christopher Krügel, and Giovanni Vigna, *Compa: Detecting compromised accounts on social networks.*, NDSS, The Internet Society, 2013.
10. Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis, and Evangelos P. Markatos, *Detecting social network profile cloning.*, PerCom Workshops, IEEE, 2011, pp. 295–300.
11. Kyumin Lee, James Caverlee, and Steve Webb, *Uncovering social spammers: social honeypots + machine learning.*, SIGIR (Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Efthimis N. Efthimiadis, and Jacques Savoy, eds.), ACM, 2010, pp. 435–442.
12. Juan Martinez-Romo and Lourdes Araujo, *Detecting malicious tweets in trending topics using a statistical analysis of language.*, Expert Syst. Appl. **40** (2013), no. 8, 2992–3000.
13. Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna, *Detecting spammers on social networks*, ACSAC (Carrie Gates, Michael Franz, and John P. McDermott, eds.), ACM, 2010, pp. 1–9.
14. Chao Yang, Robert Chandler Harkreader, and Guofei Gu, *Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers*, Recent Advances in Intrusion Detection (Robin Sommer, Davide Balzarotti, and Gregor Maier, eds.), Lecture Notes in Computer Science, vol. 6961, Springer Berlin Heidelberg, 2011, pp. 318–337.