Twitter Games: How Successful Spammers Pick Targets

Vasumathi Sridharan, Vaibhav Shankar, Minaxi Gupta
School of Informatics and Computing, Indiana University
{vsridhar, vshankar, minaxi}@cs.indiana.edu *

ABSTRACT

Online social networks, such as Twitter, have soared in popularity and in turn have become attractive targets of spam. In fact, spammers have evolved their strategies to stay ahead of Twitter's anti-spam measures in this short period of time. In this paper, we investigate the strategies Twitter spammers employ to reach relevant target audiences. Due to their targeted approaches to send spam, we see evidence of a large number of the spam accounts forming relationships with other Twitter users, thereby becoming deeply embedded in the social network.

We analyze nearly 20 million tweets from about 7 million Twitter accounts over a period of five days. We identify a set of 14,230 spam accounts that manage to live longer than the other 73% of other spam accounts in our data set. We characterize their behavior, types of tweets they use, and how they target their audience. We find that though spam campaigns changed little from a recent work by Thomas et al., spammer strategies evolved much in the same short time span, causing us to sometimes find contradictory spammer behavior from what was noted in Thomas et al.'s work. Specifically, we identify four major strategies used by 2/3rd of the spammers in our data. The most popular of these was one where spammers targeted their own followers. The availability of various kinds of services that help garner followers only increases the popularity of this strategy. The evolution in spammer strategies we observed in our work suggests that studies like ours should be undertaken frequently to keep up with spammer evolution.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues - *Abuse and Crime Involving Computers*

*Sridharan and Shankar participated in this work when they were graduate students at Indiana University.

ACSAC '12 Dec. 3-7, 2012, Orlando, Florida USA

General Terms

Security, Measurement

Keywords

Spam, Twitter, Online social networks (OSNs)

1. INTRODUCTION

Email spam has been a problem for decades. As email spam filtering programs have improved, with many claiming 99% or higher accuracies, spammers have looked for other avenues. Online social networks (OSNs), such as Twitter, are obvious new targets because of their sharp rise in popularity. Twitter alone boasted 140 million users as of March 2012 [20]. Fighting spam on OSNs requires new types of filtering techniques, owing to the fundamental differences between email and OSNs as communication media. As an example, in the case of Twitter, the unit of communication, *tweet*, is merely 140 characters, limiting the efficiency of traditional spam filters which rely on a certain size of text to be effective. Owing to the differences, both operators of OSNs as well as the research community have recently been actively pursuing the topic of spam on OSNs.

An over-arching theme across recent spam-related research work on Twitter has been the characterization of spam activity and building classifiers to aid in the identification of spammers and spam tweets. Our work complements existing research around this theme by examining an untouched aspect of Twitter spam, which is that we do not know how spammers pick their targets. Having this knowledge can complement existing efforts to defeat spam by protecting the targets and also by identifying spammers. Interestingly, this investigation is possible on OSNs because they are contained ecosystems where both spammers and the targets being spammed reside as users within the same system. This is in sharp contrast to traditional email spam where email addresses sending and receiving spam belong to different administrative entities.

In our work, we examine over 80K spam accounts on Twitter and investigate their targets. The strategies we bin spammers under are motivated by a combination of the functionality provided by popular software known to automate spam on Twitter and a tstudy of the behavior of worst offenders. The key contributions of our work are along the following two dimensions:

- **Strategies for picking targets:** We find five key strategies being used by Twitter spammers. We find that

Twitter spammers overwhelmingly target their own followers, who are expected to be voluntary subscribers of spammers' tweets. This finding is in sharp contrast to the observation made by Thomas et al. [16] just about a year ago where they found that spammers often failed to garner followers. A plausible explanation for the difference in observation is that spammers have evolved to adapt around Twitter's effort to fight spam [19]. This view is supported by observations about the thriving underground economy around buying Twitter followers [13], which makes it easy for spammers to adapt in the manner we observed.

In terms of other strategies used to find targets, we find that a relatively smaller number of spammers also target the followers of other popular accounts and even search for targets whose tweets contain keywords of interest to their spam campaigns. Finally, an even smaller number of spammers hijack popular discussion topics (referred to as *trending topics*, in Twitter parlance) to increase the possibility that their tweets are found by a large number of users interested in following those topics. The knowledge about these strategies can be used to actively identify spammers and to protect their targets from being spammed.

- **Observations about spammer behavior:** During the course of our work, we made multiple observations about spammer behavior that can be used as features to improve the performance of existing spam classifiers for Twitter. First, almost 3/4th of the spammers use a certain tweet type exclusively when only 13% of good users on Twitter have this behavior. In fact, 2/3rd of the spammers target only their own followers through the spam tweets. In contrast, only 10% of good Twitter users target only their own followers. We also learnt that the method used for posting tweets can also predict whether the tweet is spam or not, primarily because spammers seem to prefer different modes of posting their tweets than other Twitter users.

The key motivation for our work was to complement existing spam defenses and we believe that the insights gained can effectively serve that purpose. Additionally, we were able to draw comparisons with Thomas et al.'s work done a year ago, where they examined the behavior of Twitter spammers. The comparisons revealed that even though the spam campaigns have remained the same, Twitter spammers have evolved in their strategies in a mere one-year period in an attempt to stay afloat amidst Twitter's anti-spam efforts. This observation suggests that strategies of spammers need to be re-examined frequently in order to stay ahead of the game.

The rest of this paper is organized as follows. In Section 2, we discuss the methodology for our data collection and how we differentiate successful from unsuccessful spammers in our data set. Section 3 provides background information about types of tweets on Twitter and discusses our observations on the types used by spam profiles. In Section 4, we describe the major strategies we found which are used by spammers to find their target users. In Section 5, we provide our observations regarding applications used by spammers. Section 6 provides a discussion on some interesting observations in our data set. Sections 7 and 8 provide related work and conclusions, respectively.

2. DATA COLLECTION AND OVERVIEW

We used Twitter's streaming API to collect tweets for November 1st, 2011. The API samples one in ten tweets and makes them available to us. Our data set contained 19,991,050 tweets and 7,078,643 unique Twitter user account profiles. Since we needed a set of spam profiles to test, we used the output of Twitter's suspension policy as ground truth. To determine which accounts were suspended by Twitter, we visited `http://www.twitter.com/<username>` for each username contained in the profiles in our data set and looked for a string indicating suspension of the account. This process yielded 82,274 suspended profiles. These profiles are similar in nature to those analyzed by Thomas et al. in there recent work where they analyzed suspended Twitter accounts. While we analyze this data for comparison and to see how Twitter spammers have evolved in the last one year since Thomas et al.'s work, we further applied a couple of filtering criteria in order to analyze spammer strategies. First, we eliminated profiles that tweeted in languages other than English since otherwise we would not be able to analyze their tweets against various spamming strategies. This reduced our dataset to 53,083 profiles. Further, to run each profile against various spamming strategies, we needed a certain number of tweets from each. We set a threshold of 10 tweets within five days of the data collection day as our threshold and eliminated profiles that tweeted less number of tweets within the 5-day period. The choice of a 5-day period was guided by the intuition that a profile taking much longer than 5 days to tweet ten times is either inactive during our data collection period or was suspended for reasons other the sending high volume spam. This step pruned our data set to 14,230 profiles. We refer to these profiles as *successful spam profiles* subsequently in this paper and the rest of the suspended profiles as *unsuccessful spam profiles*.

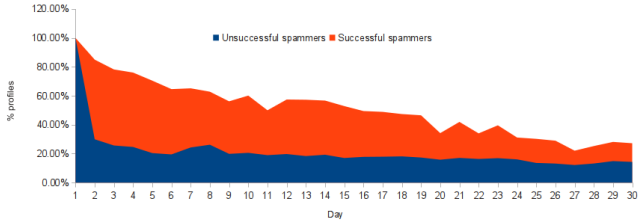

Figure 1: Lifetime of successful and unsuccessful spam profiles over one month

In the 5-day period, the entire group of 82,274 suspended profiles tweeted 1,353,340 times, of which 2/3rd of the tweets were from successful spam profiles, which comprised only 17% of the profiles. In fact, almost half of the tweets from unsuccessful spam profiles arrived on the first day of our data collection. These numbers indicate that most of the activity from unsuccessful spam profiles is front loaded, perhaps because they get suspended quickly. Figure 1 shows the percentage of successful and unsuccessful spam profiles that survive on each day for the month of November, 2011. Since we do not know how long any profile was alive before November 1st or how long it lived after November 30th, the lifetimes in this Figure are lower bounds, in that we underestimate the lifetime of all profiles. This Figure shows

Type	Successful spammers		Unsuccessful spammers		Other users	
	% Tweets	% Profiles	% Tweets	% Profiles	% Tweets	% Profiles
Regular	89.1%	91.2%	54.5%	78.7%	41%	93%
Replies	4.7%	20.9%	21.6%	32.8%	37%	80%
Mentions	5.1%	20.1%	18.4%	24.9%	11%	50%
Retweets	1.1%	8.6%	5.5%	9.5%	8%	49%
Total tweets	917,258		436,047		1,919	
Total profiles	14,230		68,044		100	

Table 1: Tweet types of successful and unsuccessful spam profiles and of regular Twitter users

that 70% of unsuccessful spam profiles and 15% of successful spam profiles get suspended on the first day. Work done by Thomas et al. a year ago [16] looked at all spammers' collectively and found that 77% of spam profiles were suspended on the first day and 92% within three days. We suspect that since unsuccessful spam profiles are a larger fraction of profiles in our data, that they dominated their data set as well, potentially masking the behavior of successful spam profiles.

3. TWEET TYPES

Twitter users can send various types of tweets. A basic tweet, referred to as a *regular* tweet in the rest of this paper, is received by all followers of a sender. It appears both on sender's *home timeline* as well as the *home timeline* of each of the sender's followers.

Three other kinds of tweets are allowed in Twitter. The first is called a *reply* tweet, which is sent as a reply to a tweet. In addition to appearing on the *home timeline* of the receiver provided they are following the sender, a reply tweet appears on the *home timeline* of anyone following both the sender and the recipient. The second type of tweet is a *mention* tweet, which is a tweet directed at a specific Twitter user. It is much like a reply tweet except that it is more broadly visible since it appears on the *home timeline* of anyone who is following just the sender. Finally, the last type of tweet is a *retweet*, which is a mechanism to forward tweets to one's followers. This is visible on the *home timeline* of each of the sender's followers.

Note that each user's *home timeline* is private to the user, implying that the tweets appearing there cannot be seen by anyone else. In turn, this implies that no one, including spammers, can misuse any information contained in the tweets, such as recipients or tweet content. However, tweets sent by each user also appear on his/her *profile timeline*, which is public by default. Tweets contained there can be searched and exploited by spammers.

Table 1 shows the tweet types seen in our data. In order to compare spammer behavior with that of other users, we randomly picked 100 regular profiles in our data. We see that an overwhelming number of profiles across both spam and other users used regular tweets. *Successful spam profiles made the heaviest use of regular tweets, with 89% of their tweets being regular.* In comparison, half or less of the tweets from unsuccessful spam profiles and other users were regular. Reply tweets, followed by mentions, were also popular among unsuccessful spam profiles as well as regular users. Successful spam profiles made relatively less use of these types of tweets, perhaps because Twitter is known to suspend accounts which send large numbers of replies or mentions [19].

Retweets were the least popular for both types of spam profiles, with successful ones using them even less than the unsuccessful ones. Even though retweets were the least popular among other users as well, they still commanded a significant fraction of their tweets. This finding is intuitive since spammers get very little leverage out of retweets, which are similar to unedited forwards of email messages but without any control over the target audience since retweets only go to one's followers.

We note that work done by Thomas et al. a year ago [16] found that 52% of spam profiles made use of mention tweets. The corresponding fractions for successful and unsuccessful spam profiles in our data are 1/5th and 1/4th respectively. For other users, it is 50%. Since their data set was sampled similar to ours, we conclude that Twitter spammers have evolved their strategies in the last one year. Finally, when looking at what percentage of profiles made use of a specific type of tweet, Table 1 suggests that spammers are using fewer types of tweets compared to other users overall. We investigate this issue in detail next.

Table 2 shows the percentage of profiles that exclusively used one type of tweet. We find that over *3/4th of successful spam profiles exclusively used only one type of tweet when the corresponding percentage for unsuccessful spam profiles is 2/3rd and for other users is 14%.* This observation suggests that spammers focus on a limited set of spamming strategies while regular users use various tweet combinations. Further, 2/3rd of successful spam profiles exclusively use regular tweets, when the corresponding percentage for unsuccessful spam profiles is about half and that for good profiles is 11% (see Table 2). Both of these characteristics could be used as features in identifying spam profiles.

Tweet type	Successful spammers	Unsuccessful spammers	Other users
Exclusively regular	68.3%	49.3%	11%
Exclusively replies	4.5%	6.8%	2%
Exclusively mentions	2.4%	10.5%	1%
Exclusively retweets	0.28%	1.0%	0%

Table 2: Spammer profiles using a single type of tweet exclusively

4. STRATEGIES FOR PICKING TARGETS

In the following sections, we describe the different strategies employed by successful spammers in finding their targets.

4.1 Spamming Ones Own Followers

We saw in Section 2 that over 2/3rd of successful spam profiles exclusively use regular tweets to send spam. Clearly,

```
http://www.amazon.com/Duke-Blue-Devils-Foot-Bean/dp/B000UIKZKC?SubscriptionId=AKIAJ7TIJCJQNX6EL3PQ&tag=
scot0e-20&linkCode=sp1&camp=2025&creative=165953&creativeASIN=B000UIKZKC&utm_source=twitterfeed&utm_
medium=twitter
```

Figure 2: An example Amazon affiliate program link (affiliate ID=scot0e-20)

for their spam campaigns to succeed, spam profiles targeting only their own followers need to have followers. (A discussion of the methods spammers employ to find their followers is in Section 6.2.)

We begin by examining the followers of all spam profiles present on the first day of our data collection. Figure 3 shows the number of followers for both successful and unsuccessful spam profiles. We note that about 1/3rd of successful spam profiles have over a 100 followers. In contrast, only 1/6th of unsuccessful spam profiles reach that number. On the other hand, only 5% of successful spam profiles have zero followers and about 1/3rd have less than 10. In comparison, nearly 40% of unsuccessful spam profiles have zero followers and a total of 2/3rd have less than 10. Thomas et al. noted in their work that 89% of spam profiles have less than 10 followers. The contrast with our observations makes us believe that not only have all spammers become smarter about acquiring followers but also that the successful spammers fare better in their follower counts. Focusing on the follower count of the group of 2/3rd successful spam profiles that exclusively used regular tweets (Figure 3), we note that the *spam profiles that use exclusively regular tweets maintain a large number of followers, with almost 1/3rd of them having 100 or more followers.*

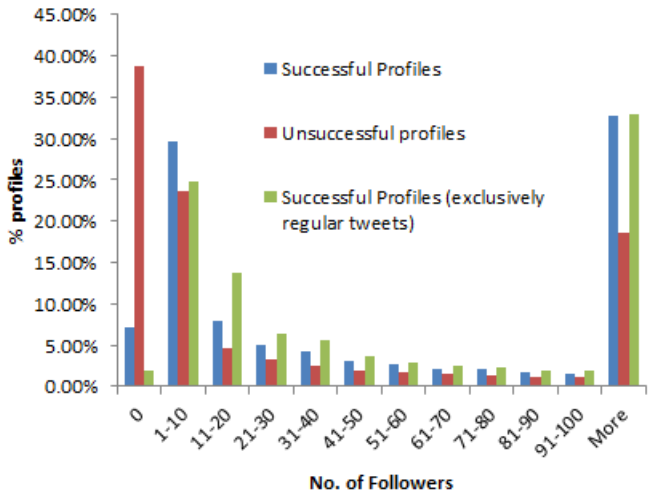

Figure 3: Follower counts

4.1.1 Spam Campaigns

To gain insights into spam campaigns of spammers using this strategy, we examine the regular tweets of all the 14,230 successful spam profiles. We only consider ones for which we have at least 10 regular tweets with links in our data. The cut-off ensures that we have sufficient tweets to judge the nature of their campaigns. This cut off left us with 7,704 spam profiles. In order to study the destinations of their spam campaigns, we require that 80% of the final destination links (upon traversing all HTTP redirects for each URL) for each profile lead to the same domain name. This heuristic ensures that we capture where these links lead while allowing for host-name and directory-level variations in URLs, which spammers often leverage to create distinct-looking links.

A total of 6,630 (86% of those with 10 links in our data) spam profiles met our criterion and were leading their followers to 559 different domains. Two of the domains were particularly noteworthy. In the first, the final destination was `t.co`, which is Twitter's own URL shortener. Since URL shorteners are never the destination page, we looked up the destination pages manually and found that they were Twitter's warning pages, informing the visitor that the link in the tweet was leading a malware-serving site. (The user still has the option of clicking on the original destination.) A total of 1,822 profiles led to a `t.co` warning page and clearly contained one or more spam campaigns leading to malware.

The second interesting domain was where 1,741 profiles were acting as *Amazon affiliates* because they were each leading their targets to `Amazon.com` URLs similar to the one shown in Figure 2. Thomas et al.'s work also found a large number of affiliate spam. However, the spam tweets in their data used mention tweets, while we find spam profiles to be using regular tweets instead, indicating a shift in strategy. Almost all of the spam URLs had `amzn.to` as the starting point landing page in tweets, which is `bit.ly`'s specialized URL shortening service for Amazon. `Amazon.com` accounted for over 55% of the 917,294 tweets by successful spam profiles, and contained a total of 76 unique Amazon affiliate IDs.

In order to understand the different `Amazon.com` campaigns, we examined the affiliate IDs and the spam profiles associated with them. All profiles using the same affiliate ID were clearly part of the same campaign. Common profiles across multiple affiliate IDs also implied that they belonged to a spam campaign. We found 19 `Amazon.com` campaigns in our data set where the spammers were using regular tweets to target their followers. Table 3 shows the top five campaigns where a spammer either used a large number of Twitter profiles or affiliate IDs. The largest of these involved over 1.5K Twitter profiles and sent almost half a million tweets over our data collection period.

Amazon campaign	Twitter profiles	Affiliate IDs	Tweets
1	1,519	55	446,552
2	163	1	46,848
3	18	1	4,853
4	8	4	4,441
5	3	1	43
Unique spam profiles across all 19 campaigns		1741	

Table 3: Top-5 Amazon affiliate spam campaigns where spammers used regular tweets to target their own followers

4.1.2 Spam Profiles Binned

Of the 14,230 successful spam profiles, 92.1% (12,979) used regular tweets and 68.3% (9,723) used regular tweets exclusively. The analysis in this section suggests that 6,630 spam profiles were targeting their own followers in their spam campaigns. It is noteworthy that of these, 93.6% (6,208) were using regular tweets exclusively.

4.2 Spamming Followers of Popular Profiles

As an alternative to targeting ones own followers, spammers can find suitable targets by simply targeting other profiles' followers. This is advantageous in cases such as when, say, a spammer wants to target music lovers. He/she can send spam to the followers of music celebrities, leveraging the fact that many of the celebrities have already acquired a large number of followers. Targeting the followers of popular profiles is in general an easy strategy a spammer can use, even if the goal is not to target a specific type of users. In fact, softwares such as TweetAttacks [18] and TweetAdder [17] have readily available automation spammers could exploit to target followers of any profile. Note that the tweets sent using this strategy will have to make use of reply or mention tweets since otherwise, they will only go to the profile's followers.

A total of 4,086 (28.7%) spam profiles made use of reply or mention tweets in our data. In order to determine how many of them targeted followers of other profiles, we needed to have at least a few samples of the users who received their spam. We used a threshold of four unique Twitter users receiving spam from a specific spam profile and checked if at least 50% of the targets for each spammer were following the same profile. If so, the spammer likely targeted the followers of that profile.

Of the spammers that made use of reply or mention tweets, 3,528 (86.3%) had targeted at least four different users in their spam. For each spammer's targets, we collected information about all profiles each target was following using the Twitter API. However, since API calls are expensive, we collected only the first 5000 followings for each target. When checking if at least 50% of the targets for each were following a specific profile, we found that 877 profiles fit the heuristic. They contributed an average of just under 23 reply or mention tweets each, which is much lower than the average of 64 tweets an average spam profile in our data contributed (see Table 1). This suggests that the volume of spam from this strategy is not high. Examining the tweets of this group of 877 spam profiles, only 272 had more than 10 reply or mention tweets with links. Of these, 225 profiles had more than 80% of their links going to the same domain, implying that these profiles were pointing their targets to a specific domain and potentially running spam campaigns.

The most popular of these domains was `hellojb.info`, which made up 18% (157) of the total 877 profiles and 18% of the total reply or mention tweets from this group of spammers. 149 of the profiles from this campaign used reply tweets to target the followers of celebrity Justin Beiber with the link that claimed to have some information on how to get Justin Beiber to follow them.

The second most interesting domain was where six spam profiles were acting as Amazon affiliates. Their tweets carried four unique affiliate IDs, suggesting that there are up to four spam campaigns here (see Table 4). Though of a much smaller scale, these campaigns are of a similar nature to the Amazon affiliates campaigns found in Section 4.1. However, since none of the profiles or affiliate IDs here are common with those found in Section 4.1, we believe that the spammers behind these campaigns are a different group from those found in Section 4.1.

Amazon campaign	Twitter profiles	Affiliate IDs	Tweets
1	3	1	34
2	1	1	14
3	1	1	11
4	1	1	11
Unique spam profiles across all 4 campaigns		6	

Table 4: Amazon affiliate spam campaigns where spammers used reply or mention tweets to target others' followers

4.2.1 Spam Profiles Binned

A total of 877 spam profiles were binned under this strategy. Of these, 26 were common with the 6,630 binned under Section 4.1. These were clearly targeting their own followers through regular tweets and others' followers through reply or mention tweets. Overall, thus far, we have binned 7,481 (52.5%) of the 14,230 profiles of successful spammers.

4.3 Spamming based on Keywords in Tweets

Spammers can also pick their targets based on the content of tweets from Twitter users. For example, a spammer who has a campaign revolving around phones may target users who have used the term "phone" in their tweets. Finding such targets is easy for spammers since Twitter already facilitates searching tweets based on keywords. In fact, that softwares such as TweetAdder already support automating this spamming strategy suggests that some spammers may be making use of it.

Spammers using this strategy could use reply tweets or mention tweets to send spam to their chosen targets. However, looking at a mention tweet, one cannot determine which specific tweet triggered a response from the spammer. This rules out the scrutiny of 1,408 spam profiles that used mention tweets but not reply tweets. Fortunately, it is possible to check for the use of this strategy when reply tweets are used, since reply tweets contain an identifier, *status ID*, for the original tweets that were replied to.

2,969 successful spammers in our data (21% of 14,230) made use of reply tweets, some in conjunction with other types of tweets. In order to judge if a spam tweet contained a keyword, we first had to identify possible keywords. Toward this goal, we used the term frequency-inverse document frequency (TF-IDF [8]) statistic which reflects how important a word is to a document in a collection. Upon computing the TF-IDF score of about 7 million words present in the tweets from successful spammers, we picked the top 50K as possible keywords. Then for each reply tweet of each spam profile that had at least three reply tweets, we extracted the source tweets (tweets they were replied to) by using the status ID present in the tweets and querying the Twitter API. In the next step, we looked for common keywords in source tweets for the each spam profile. If a word appeared in over a percentage threshold of a spam profile's tweets and was one of the 50K words we chose as keywords, we took it to imply that the spammer targeted authors of tweets with

specific keywords. Since we had varying number of source tweets for each spammer, we used four slabs to judge if a spam profile was using this strategy to find targets. For the cases where we had exactly 3 tweets, it was important to set a very high threshold. So, we set a threshold of 100%, implying all 3 source tweets had to have the keyword in question. For profiles with more than three but less than 10 tweets, we required that at least 50% of the tweets had to have the keyword. Further, for cases when we had more than 10 but less than 20 tweets, we relaxed the threshold to 40%. Finally, when more than 20 tweets were available, we required that at least 30% should have the keyword.

Of the 2,969 spam profiles that used reply tweets, 2,419 (81.4%) had at least three reply tweets each. Of these, 1,004 (41.5%) passed the heuristic to fit this strategy. Further, 710 of these profiles had at least 10 tweets with links and at least 80% of their links were going to a single domain, implying that these spam profiles were likely running campaigns. The most popular domain was `jbet.info`, which involved 173 profiles that were promoting this health-related website using keywords "breakfast", "lunch", and "dinner" to locate targets. The next most popular, `hellojb.info`, was also caught in Section 4.2 since the spammer was targeting the followers of celebrity Justin Bieber. It was caught here due to the presence of keyword "justinbieber" in the tweets of targets and the profiles involved in both cases were the same set. Finally, the Amazon affiliate marketing campaign featured prominently in this category as well. We saw six different campaigns using a diverse range of keywords, ranging from "buy", "iPhone", "weight", "aging", and "jailbreak" (see Table 5). The largest two campaigns involving 73 and 19 spam profiles respectively, had three and one profiles common with Amazon campaigns found in Section 4.2. Also, campaign 4, involving three spam profiles, shared a profile with Section 4.2. A closer inspection of the common profiles suggests that the spammers simply targeted keywords and got followers of specific popular profiles as a result.

Amazon campaign	Twitter profiles	Affiliate IDs	Tweets
1	73	1	1,059
2	19	1	245
3	5	2	66
4	3	1	30
5	1	1	15
6	1	1	13
Unique profiles across all 6 campaigns		102	

Table 5: Amazon affiliate spam campaigns where spammers used keywords to find targets

4.3.1 Spam Profiles Binned

A total of 1,004 spam profiles were binned under this strategy. Of these, only one was found in Section 4.1 while 210 found in Section 4.2. Of the latter, 150 were simply due to the `hellojb.info` campaign discussed earlier. Overall, thus far, we have binned 8,274 (58.1%) of the 14,230 profiles of successful spammers.

4.4 Trending Topics Hijacking

Twitter users can tag their tweets with a topic by using a keyword after the symbol '#'. This keyword is referred to as a hashtag in Twitter parlance. Popular hashtags become trending topics, which attract visibility. Spammers have been known to hijack trending topics to increase the visibility of their spam campaigns [16]. Here, we examine how many of the successful spammers in our data set are exploiting this strategy. Note that spammers can make use of this strategy in conjunction with other spamming strategies, such as spamming their own followers or followers of popular profiles.

A spammer using this strategy may use any of the various types of tweets, regular, mentions, or replies so long as the tweet contains a hashtag. We analyzed the 4,327 spam profiles in our data set that had at least one hashtag in their tweets. In order to identify spammers who tried to hijack trending topics, we pre-calculated a set of 200 most popular hashtags out of the total 466,597 hashtags present in the tweets in our data set. Next, just like in Section 4.3, we required that each spammer profile have a minimum of three tweets with hashtags. This ruled out 824 profiles leaving us with 3,503 (81%) profiles to examine. On each profile, we applied the same slabs as we applied to determine keywords in Section 4.3 to determine whether or not it abused a popular hashtag for spam distribution.

Our heuristic concluded that 1,043 spam profiles hijacked trending topics. Of these, 174 likely ran a campaign since they led their targets to a specific domain as they had at least 10 hashtag tweets with links and 80% of their tweets pointed to a particular domain. The top campaigns used fewer profiles and proportionally fewer tweets compared to those found in Section 4.3 and were promoting politically-oriented, gaming-related and various kinds of unsavory websites. Two domains in these campaigns were the most interesting. The first was the Amazon affiliate ID campaign, which was found in all previous sections. Here, we found five spam campaigns (shown in Table 6). All but one campaign was already identified in Section 4.1, indicating that spammers simply tried to cast their net wide. The second interesting domain was `adf.ly`, where 19 spam profiles were hijacking trending topics to advertise this ad-based shortener. Such shorteners were observed by Thomas et al. also. The revenue model for these shorteners is to show advertisements before leading the visitor to the destination link.

Amazon campaign	Twitter profiles	Affiliate IDs	Tweets
1	5	5	836
2	1	1	471
3	1	1	40
4	1	1	31
5	1	1	10
Unique profiles across all 5 campaigns		9	

Table 6: Amazon affiliate spam campaigns where spammers hijacked trending topics to gain attention

4.4.1 Spam Profiles Binned

Given that this strategy merely requires the addition of an extra hashtag to a spam tweet, it can be easily combined with other strategies. Hence, we see significant commonality between profiles binned under this strategy and others discussed before. Specifically, we found that 523 of the 1,044 profiles binned under this strategy were using regular tweets to spam their followers (discussed in 4.1). Another 14 profiles were common with Section 4.2. Finally, three profiles

were common with those in Section 4.3. Overall, this strategy binned 503 new spam profiles not binned through previous strategies, bringing the total binned profiles to 8,777 (61.7%) of the 14,230 profiles of successful spammers.

4.5 Targeting Own Followers by Retweets

We saw in Section 4.1 that regular tweets are a popular way for spammers to target their followers. Retweets can be used as an alternative because they also reach ones followers, except that they are restrictive because the tweet content has to be borrowed verbatim from another tweet and cannot be modified. However, according to the TweetAttacks software, retweets have a higher click rate than normal tweets, which may make spammers prefer them. In fact, TweetAttacks referred to the strategy of using retweets as "retweet attacks".

Retweets were the least popular among spammers as well as other users, with successful spammers using them even lesser than others (see Table 1). 1,230 successful spam profiles used retweets at all. Of these, only 70 had greater than 10 retweets with links and only 28 had 80% of their links pointing to a single domain, suggesting that they were running campaigns. Incidentally, each of these 28 were using retweets exclusively. 26 of them were retweeting content from a Twitter profile called *omgwire*. The profile seemed to be promoting a celebrity gossip website, `omgwire.com`, so the retweets were tailored to that promotion. The other two domains were also retweeting content from two separate Twitter profiles and promoting a website each.

4.5.1 Spam Profiles Binned

This strategy only helped bin 28 additional spam profiles but none were common with those found under previous strategies. Overall, we have thus far binned 8,805 (61.9%) spam profiles of the total 14,230 profiles of successful spammers.

5. POSTING METHODOLOGY

Figure 4: Examining the "via" field at the bottom of a tweet helps determine how it was posted

Twitter allows posting tweets using a variety of methodologies, including their Web interface and various Twitter-provided and third party clients which post through the Twitter API. Examples of such clients include Tweet Button, Twitter's mobile interface, and specialized applications for various smartphone platforms, such as BlackBerry, iPhone, and Android. Additionally, there are RSS-to-tweet services such as *twitterfeed*,`dlvr.it` and Google's client for the users of their blog service,blogspot.com that automatically post a link to Twitter account each time a new content is published on a blog.

Twitter facilitates knowing how a tweet was posted. For example, the bottom of the tweet shown in Figure 4, which is one of the spam tweets from the Amazon campaign discussed throughout Section 4, shows that it was posted via Twitterfeed, a third-party client.

Looking into how successful and unsuccessful spam profiles differ in their posting behavior versus other users helped draw interesting conclusions which we present next. Table 7 shows the top six most popular ways of posting, based on the number of tweets posted and the number of profiles that make use of it. We find that *Twitterfeed is the most popular among spam profiles, with successful ones exploiting it for 2/3rd of their tweets.* The Web is popular among all three groups, with close to 40% of the profiles in each group posting via the Web. Note that popular software, such as TweetAdder and TweetAttacks, automate their posts so they appear to have been posted via the Twitter Web interface. Twitter's mobile Web interface, which is the version of Twitter.com for mobile devices, is popular among unsuccessful spammers but not among successful spammers and other users.

Also, there is evidence of spammers using fewer dedicated apps when compared to other users as shown in Table 8. The average suggests that the organic profiles use several different apps, where as spammers (successful or unsuccessful) have fewer dedicated apps.These findings indicate that the method of posting tweets itself could be used to predict if a tweet is more likely to be spam versus good.

Type of profile	No. of apps used	users per app
Other users	65	0.68
Successful	790	18.01
Unsuccessful	1,599	42.55

Table 8: Average number of apps used by a single profile under each category of users

6. DISCUSSION

In the pervious sections, we discussed how successful spam profiles find their targets and what tools are often associated with spam accounts. In this section, we discuss two additional aspects of our study. First, we look at the set of successful spam accounts which were not accounted for in any of the strategies described in Section 4. We discuss what strategies may have been missed and offer explanations based on limitations in our dataset to discover such profiles. Next, we discuss the concept of gathering followers. As was shown in 4.1, a large number of successful spammers have a significant number of followers. We briefly investigate what methods are used by spam accounts to gather followers and remark upon two major types of follower gathering schemes.

Type	Successful spammers		Unsuccessful spammers		Other users	
	% Tweets	% Profiles	% Tweets	% Profiles	% Tweets	% Profiles
twitterfeed	65.3%	25.1%	8.18%	5.4%	3.5%	4.0%
Web	13.3%	38.7%	41.5%	45%	24%	45%
Tweet Button	3.5%	15.8%	2.5%	3.7%	0.0005%	1%
Mobile Web	3.3%	8.3%	10.6%	24.7%	4.5%	9%
`dlvr.it`	3%	2.7%	1.44%	1.4%	0.46%	1%
Google	1.1%	1.7%	0.2%	0.3%	0%	0%
Total tweets	917,258		436,047		1,919	
Total profiles	14,230		68,044		100	

Table 7: Tweet posting methodology of successful and unsuccessful spam profiles and of regular Twitter users

6.1 Unbinned Spam Profiles

We discussed five strategies that successful spammers use to pick their spam targets in Section 4. Collectively, these strategies helped understand the spamming behavior of 62% of the 14,230 profiles of success spammers. In trying to understand the behavior of unbinned spam profiles, we learnt a few things. First, of the 5,425 unbinned profiles, the largest chunk, 64.7%, exclusively made use of regular tweets and could not have directly targeted any other Twitter users but their followers. Owing to this observation, they could only be binned under Section 4.1 amongst the strategies we describe. Recall that in Section 4.1, we considered a profile binned if it appeared to be running a campaign, defined by at least 10 tweets with links and 80% of the links leading to a single domain. Since alternative definitions for campaigns is possible, we tried a few different ones to see their impact on the proportion of unbinned profiles. For example, if we require that a profile only have at least five tweets with links and when the number of tweets with links is less than 10, 80% of them be leading to a specific domain and when the number of tweets with links is 10 or more, 50% of them be leading to a specific domain, we bin 8,034 profiles in Section 4.1 as against 6,630. In turn, 72% spam profiles get binned. This simple exercise suggests that further experimentation with definitions of campaign is one avenue for binning more spam profiles.

Of the remaining 1,910 unbinned profiles, 89.7% sent at least one mention tweet and 328 sent exclusively mention tweets. Note that of the two strategies involving mention tweets, described in Sections 4.2 and 4.3, the latter had to ignore mention tweets due to the unavailability of keywords spammers may have targeted from users' original tweets. This caused us to miss an opportunity of binning more spam profiles that used mention tweets.

Further, spam automators, such as TweetAdder, provide other ways spammers could use to pick their targets. For example, targets could be chosen based on their geographical location and language, which are pieces of information typically associated with Twitter profiles. However, given that the profiles we investigated were already suspended, we did not have access to information about their followers. Such information may have helped us investigate more strategies used, especially by regular tweeting profiles, to identify targets.

6.2 Garnering Followers

Given that about 90% of successful spam profiles make use of regular tweets and 2/3rd of them use them exclusively, how spammers garner followers is an interesting consideration. Indeed, Figure 3 confirmed that a large number of successful spammers have a respectable number of followers.

There are many ways spammers can use to garner followers. A popular mechanism is to become a part of one or more peer-driven communities that encourage following back. We discovered two interesting communities in our data, `#InstantFollowBack` (#IFB) and `#TeamFollowBack` (#TFB).

The first of these is controlled by a third-party client under the Twitter profile name 'instantfollowBA' that allows a Twitter user to find and follow users on Twitter by requiring them to be listed under it's follow back community #InstantFollowBackGradeA. The requirements for this is to have public account with minimum 500 followers and follow back 99% of the them on a every day basis which the client app claims to make it easier by providing a few tools. Additionally, it rewards each user with points that offer different levels of promotions. Each user is given a 'status' that indicates the number the 'Gold' and 'Experience' points earned by them. Users can increase their points by being a part of #InstantFollowBackGradeA and by tweeting (advertising) about 'instantfollowBA'. As the level of a user increases, more points are awarded. After collecting a sufficient number of points, a user gains center stage in promotions, such as retweets to all of the main #IFB profile's followers, retweets by others in the community to increase the profile's visibility in Twitter, display of banner ads by several members of the #IFB community and so on. This method of incentivizing users to join a follow back community is unique though its popularity may be limited due to the complicated nature of the setup. In fact, we found only 217 profiles from our data set that were involved in this scheme. The second community in our data, #TFB, is a Twitter hashtag used by numerous follow back groups with goals similar to #IFB. Profiles involved in these schemes sign their tweets with the #TFB hashtag. Often, they add the hashtag to their publicly searchable profile information. This makes it easy for anyone requiring followers to find profiles willing to follow them back without regard to who is requesting or what content is being tweeted. We found 509 profiles in our data using the #TFB hashtag, all in Section 4.4.

Yet another option for increasing the number of followers is to buy them. Owing to the importance of collecting followers, websites such as, `http://getfollowrsontwitter.org/`, have cropped up that allow one to buy followers, often in increments of thousands. At the same time, online marketplaces for services, such as `fiverr.com`, routinely feature services where offerers guarantee thousands of Twitter followers for as little as $5 (see Figure 5 for an example of such an offering). Works in [6, 13] studied Twitter account mar-

kets and confirmed that buying followers is indeed prevalent on Twitter.

Figure 5: An advertisement offering to gather Twitter followers

It is important to note that the techniques described thus far only go as far in finding followers. Specifically, they are unlikely to work well for spam profiles that require relevant followers, such as in the case of exclusive regular tweeters. As described before, these are the largest chunk of successful spammers and require more intelligent strategies for garnering followers. To garner such followers, spammers could use the strategies we described in Sections 4.2 and 4.3 to locate targets and friend them. Additionally, a spammer may use publicly searchable profile information to target people based on location, interests etc. Since a good number of these targets may be genuinely interested in the content promoted by a spammer, it may increase the likelihood of them actually following the spammer back. While we have seen evidence of such activity in a few spam profiles that are alive, it is in general difficult to prove if the followers of a profile were gathered using a certain strategy. In fact, spamming only one's followers thins the line between genuine and spam content since the only real violation by spam accounts tweeting only to their followers is the use of an automated tool to send tweet in large volumes. In contrast, the dominant spammers in Thomas et al.'s work were blatant violators of Twitter terms and conditions since they largely made use of unsolicited mentions to promote spam tweets.

7. RELATED WORK

Spam on online social networks has been analyzed in a number of different ways. Benevenuto et al. [2] analyze online video spam on Youtube and employ machine learning techniques to identify spammers on YouTube. The study by Gao et al. [5] involves detecting and characterizing spam campaigns on Facebook. Here, we focus primarily on works related to Twitter, focusing particularly on previous research which is most related to our work and influenced our investigation.

Much of the research on Twitter spam has focused on building classifiers that distinguish spam profiles from non-spam profiles or spam tweets from non-spam tweets. Lee et al. [10] and Stringhini et al. [14] investigated spam on Twitter using *social honey pots*, which are profiles created specially for the purpose of being spammed and proposed a machine learning based approach to classify profiles that send spam to these accounts. Lee et al [9] also studied collective spam on Twitter by analyzing how cyber criminals exploit *trending topics* to propagate spam and devised a machine learning based methodology to detect them. Yang et al. [22] studied tactics used by spammers and employed machine learning features to detect them. Hongyu et al. [4] gathered spam messages into campaigns and used supervised machine learning framework for filtering them. Works in [1, 3, 21] studied spam by manually labeling their data sets into spam and non-spam accounts and built a classifier using account-based, content-based and behavior-based features. Classifiers of nature similar to these works can benefit from using features we found common across spam profiles in our data.

Significant research has also been done in detection and characterization of suspicious URLs in Twitter. Lee et al. [11] proposed a machine learning classifier to detect suspicious URLs by identifying characteristics of URLs in spam tweets. Thomas et al. [15] proposed a real time URL classifier that extracts features from page content, hosting infrastructure and lexical properties of URLs to detect spam urls in a Twitter stream.

Grier et al. [7] characterized Twitter spam by discussing various topics like type of tweets spammers use, click through rates of shortened URLs, types of spam accounts, blacklist performance and the techniques used by spammers to garner a wider audience.According to their analysis, 70% of spam tweets had hashtags (compared to 7.3% in our data), 11% were retweets (compared to 1.1% in our data) and 10.6% (compared to 5.1% in our data) were mentions.

Yang et al. [23] studied how spammers get embedded deeper in social networks and found three categories of users that form closer relationships with spammers that post malicious links. Jonghyuk et al. [12] proposed a spam filtering technique by analyzing sender-receiver relationship between users.

A very closely related work to our work here was done by Thomson et al. [16]. The authors collect a large number of suspended Twitter accounts over a period of time and analyze their behavior. Although we have a similar data collection methodology (albeit for a shorter period of time), we show how much Twitter spam has evolved since their study, likely to counter Twitter's current anti-spam efforts. Specifically, we find that multiple characteristics of spam accounts, such as social relationships formed, longevity and type of tweets used differ significantly in our study even though the nature of spam campaigns remains essentially the same. We focus on spam accounts which survive much longer than those set up by naive spammers and discuss their strategies.

8. CONCLUSION

We analyzed strategies of successful Twitter spammers in this paper, particularly as they relate to picking spam targets. A key finding of our work was that while spam campaigns on Twitter changed little, the spammers themselves evolved in a mere matter of one year since Thomas et al., perhaps to stay afloat amidst take-down efforts. The

complexity of their strategies are only likely to increase as more and more tools which simulate normal human behavior are developed. Given the shift in spammer behavior to integrate more closely into the social network, this will require constant scrutiny on the part of Twitter to discover newer and more complex strategies. We believe there is a need for spam classifiers to include social metadata such as follower metadata, keywords cloud, domains linked in tweets etc. along with the traditional signals used in classifiers today to achieve true success in identifying sophisticated spam profiles.

Acknowledgements

We thank Fil Menczer and the Center for Complex Networks and System Research (CNetS) at Indiana University for providing us access to the Twitter streaming API data through their Truthy project. The Truthy project and its infrastructure are supported by the National Science Foundation (NSF) grants CCF-1101743 and IIS-0811994 respectively.

The work in this paper is supported by the National Science Foundation (NSF) under Grant Number CNS-1018617. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

9. REFERENCES

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

[2] Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Chao Zhang, K. R. Identifying video spammers in online social networks. In *Workshop on Adversarial Information Retrieval on the Web (AirWeb), held in conjunction with the International World Wide Web (WWW) conference* (2008).

[3] Chu, Z., Gianvecchio, S., and Wang, H. Who is tweeting on Twitter: Human, bot, or cyborg? In *Annual Computer Security Applications Conference (ACSAC)* (2010).

[4] Gao, H., Chen, Y., Lee, K., Palsetia, D., and Choudhary, A. Towards online spam filtering in social networks. In *ISOC Network and Distributed System Security Symposium (NDSS)* (2012).

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

[6] Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., Ganguly, N., and Gummadi, K. P. Understanding and combating link farming in the Twitter social network. In *International Conference on World Wide Web (WWW)* (2012).

[7] Grier, C., Thomas, K., Paxson, V., and Zhang, M. @spam: the underground on 140 characters or less. In *ACM Conference on Computer and Communications Security (CCS)* (2010).

[8] Jones, K. S. A statistical interpretation of term specificity and its application in retrieval. In *Journal of Documentation, Vol. 28 Issue: 1, pp.11 - 21* (1972).

[9] Lee, K., Caverlee, J., Kamath, K. Y., and Cheng, Z. Detecting collective attention spam. In *Workshop on WebQuality, held in conjunction with International World Wide Web (WWW) conference* (2012).

[10] Lee, K., Caverlee, J., and Webb, S. Uncovering social spammers: Social honeypots + machine learning. In *ACM Special Interest Group on Information Retrieval (SIGIR) Conference* (2010).

[11] Lee, S., and Kim, J. Warningbird: Detecting suspicious URLs in twitter stream. In *ISOC Network and Distributed System Security Symposium (NDSS)* (2012).

[12] Song, J., Lee, S., and Kim, J. Spam filtering in twitter using sender-receiver relationship. In *International Symposium on Recent Advances in Intrusion Detection (RAID)* (2011).

[13] Stringhini, G., Egele, M., Kruegel, C., and Vigna, G. Poultry markets: On the underground economy of Twitter followers. In *ACM Workshop on Online Social Networks (WOSN)* (2012).

[14] Stringhini, G., Kruegel, C., and Vigna, G. Detecting spammers on social networks. In *Annual Computer Security Applications (ACSAC) conference* (2010).

[15] Thomas, K., Grier, C., Ma, J., Paxson, V., and Song, D. Design and evaluation of a real-time URL spam filtering service. In *IEEE Symposium on Security and Privacy* (2011).

[16] Thomas, K., Grier, C., Song, D., and Paxson, V. Suspended accounts in retrospect: an analysis of twitter spam. In *ACM/USENIX Internet Measurement Conference (IMC)* (2011).

[17] TweetAdder, 2012. `http://www.tweetadder.com`.

[18] TweetAttacks manual, 2012. `http://www.scribd.com/doc/59395233/Manual-Tweet-Attacks`.

[19] Twitter rules, 2012. `https://support.twitter.com/entries/18311-the-twitter-rules`.

[20] Twitter size, 2012. `http://blog.twitter.com/2012/03/twitter-turns-six.html`.

[21] Wang, A. H. Don't follow me: Spam detection in Twitter. In *International Conference on Security and Cryptography (SECRYPT)* (2010).

[22] Yang, C., Harkreader, R., and Gu, G. Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers. In *International Symposium on Recent Advances in Intrusion Detection (RAID)* (2011).

[23] Yang, C., Harkreader, R., Zhang, J., Shin, S., and Gu, G. Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on Twitter. In *International Conference on World Wide Web (WWW)* (2012).