Suspended Accounts in Retrospect: An Analysis of Twitter Spam

Kurt Thomas[†] Chris Grier[†*] Vern Paxson[†*] Dawn Song[†]
[†]University of California, Berkeley [*]International Computer Science Institute
{kthomas, grier, vern, dawnsong}@cs.berkeley.edu

ABSTRACT

In this study, we examine the abuse of online social networks at the hands of spammers through the lens of the tools, techniques, and support infrastructure they rely upon. To perform our analysis, we identify over 1.1 million accounts suspended by Twitter for disruptive activities over the course of seven months. In the process, we collect a dataset of 1.8 billion tweets, 80 million of which belong to spam accounts. We use our dataset to characterize the behavior and lifetime of spam accounts, the campaigns they execute, and the wide-spread abuse of legitimate web services such as URL shorteners and free web hosting. We also identify an emerging marketplace of illegitimate programs operated by spammers that include Twitter account sellers, ad-based URL shorteners, and spam affiliate programs that help enable underground market diversification.

Our results show that 77% of spam accounts identified by Twitter are suspended within on day of their first tweet. Because of these pressures, less than 9% of accounts form social relationships with regular Twitter users. Instead, 17% of accounts rely on hijacking trends, while 52% of accounts use unsolicited mentions to reach an audience. In spite of daily account attrition, we show how five spam campaigns controlling 145 thousand accounts combined are able to persist for months at a time, with each campaign enacting a unique spamming strategy. Surprisingly, three of these campaigns send spam directing visitors to reputable store fronts, blurring the line regarding what constitutes spam on social networks.

Categories and Subject Descriptors

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

General Terms

Security, Measurement

Keywords

Social Networks, Spam, Account Abuse

IMC'11, November 2–4, 2011, Berlin, Germany.

1. INTRODUCTION

As Twitter continues to grow in popularity, a spam marketplace has emerged that includes services selling fraudulent accounts, affiliate programs that facilitate distributing Twitter spam, as well as a cadre of spammers who execute large-scale spam campaigns despite Twitter's efforts to thwart their operations. While social network spam has garnered a great deal of attention in the past year from researchers, most of the interest has involved developing tools to detect spam. These approaches rely on URL blacklists [5, 20], passive social networking spam traps [12, 19], and even manual classification [2] to generate datasets of Twitter spam for developing a classifier that characterizes abusive behavior. These spam detection approaches however have not yet been used to analyze the tools and techniques of spammers, leaving the underground marketplace that capitalizes on Twitter largely obscure.

In this paper we characterize the illicit activities of Twitter accounts controlled by spammers and evaluate the tools and techniques that underlie the social network spam distribution chain. This infrastructure includes automatically generated accounts created for the explicit purpose of soliciting spam; the emergence of *spam-as-a-service* programs that connect Twitter account controllers to marketers selling products; and finally the techniques required to maintain large-scale spam campaigns despite Twitter's counter-efforts. To perform the study, we aggregate over 1.8 billion messages on Twitter sent by 32.9 million accounts during a seven month period from August 17, 2010 to March 4, 2011. Within this period, we identify accounts suspended by Twitter for abusive behavior, including spam, aggressive friending, and other non-spam related offenses. Manual analysis indicates that an estimated 93% of suspended accounts were in fact spammers, with the remaining 7% suspended for mimicking news services and aggressive marketing. In total, our dataset consists of over 1.1 million suspended accounts that we show to be spammers, and 80 million spam tweets from these accounts. In contrast to previous studies [5, 6], only 8% of the URLs we examine were ever caught by blacklists, and the accounts within our dataset are largely fraudulent, as opposed to compromised users. This enables us to provide a unique perspective on a subset of Twitter spammers not previously examined.

At the heart of the of the Twitter spam craft is access to hundreds of accounts capable of reaching a wide audience. We find that 77% of accounts employed by spammers are suspended within a day of their first post, and 92% of accounts within three days. The countermeasures imposed by Twitter's suspension algorithm preclude the possibility of attempting to form meaningful relationships with legitimate users, with 89% of spam accounts having fewer than 10 followers. In place of distributing messages over the social graph, we find that 52% of spam accounts turn to *unsolicited mentions*, whereby a personalized message is sent to another ac-

count despite the absence of a social relationship. Another 17% of accounts rely on embedding *hashtags* in their messages, allowing spam to garner an audience from users who view popular Twitter discussions via search and *trending topics*.

Beyond the characteristics of spam accounts, we explore five of the largest Twitter spam campaigns that range from days to months in duration, weaving together fraudulent accounts, diverse spam URLs, distinct distribution techniques, and a multitude of monetization approaches. Together, these campaigns control 145 thousand account that generate 22% of spam on Twitter. Surprisingly, three of the largest campaigns direct users to legitimate products appearing on *amazon.com* via affiliate links that generate income on a purchase, blurring the line regarding what constitutes spam. Indeed, only one of the five campaigns we analyze advertises content generally found in email spam [13], revealing a diverse group of miscreants in the underground space that go beyond email spammers.

Finally, within the amalgam of spam on Twitter, we identify an emerging market of *spam-as-a-service*. This marketplace includes affiliate programs that operate as middlemen between spammers seeking to disseminate URLs and affiliates who control hundreds of Twitter accounts. The most prominent affiliate program, called Clickbank, appeared in over 3.1 million tweets sent from 203 affiliates participating in the program. Other services include ad-based URL shorteners as well as account arbiters who sell the ability to tweet from thousands of accounts under a single service's control. Each of these services enables a diversification in the social network spam marketplace, allowing spammers to specialize exclusively in hosting content or acquiring Twitter accounts.

In summary, we frame our contributions as followers:

- We characterize the spamming tools and techniques of 1.1 million suspended Twitter accounts that sent 80 million tweets.
- We examine a number of properties pertaining to fraudulent accounts, including the formation of social relationships, account duration, and dormancy periods.
- We evaluate the wide-spread abuse of URLs, shortening services, free web hosting, and public Twitter clients by spammers.
- We provide an in-depth analysis of five of the largest spam campaigns targeting Twitter, revealing a diverse set of strategies for reaching audiences and sustaining campaigns in Twitter's hostile environment.
- We identify an emerging marketplace of social network *spam-as-a-service* and analyze its underlying infrastructure.

2. BACKGROUND

In this section, we discuss previous studies of underground markets as well as provide an overview of how spammers abuse the Twitter platform, previous approaches for detecting and measuring social network spam, and Twitter's own detection mechanism, which we rely on as a source of ground truth.

2.1 Twitter Spam

Twitter spam is a systemic problem. Unsolicited content appears throughout personal feeds, search query results, and trending topics. This content is distributed by fraudulent accounts explicitly created by spammers, compromised accounts, malicious Twitter applications, and even legitimate users posting syndicated content. We describe each of these components and how they tie into the Twitter platform.

2.1.1 Exposure to Spam

User Timeline: Users receive content in the form of a *user timeline* that includes tweets broadcast by each of a user's friends as well as posts that *@mention*, or tag, a timeline's owner. Spammers can inject content into this stream by either enticing a user into *following* the spammer (forming a directed social relationship), or by mentioning the user in a post. Mentions require no prior relationship to exist between a spammer and a user.

Trending Topics: Rather than forming relationships with users or targeting single users, spammers can post tweets that contain popular keywords from *trending topics*. Trends result from spontaneous coordination between Twitter users as well as from breaking news stories. Users that explore these trends will receive a feed of legitimate tweets interspersed with spam.

Search: Twitter provides a tool for searching public tweets beyond popular topics. Spammers can embed popular search terms into their tweets, similar to hijacking trends and search engine optimization. However, as there are no public signals to determine what spam tweets were accessed due to search queries, search spam is beyond the scope of this study.

Direct Messages: Direct messages are private tweets sent between two users, effectively duplicating the functionality of email. Spammers can send unsolicited direct messages to users they *follow*; users do not need to reciprocate the relationship in order to receive spam content. As direct messages are private, they are beyond the scope of this study.

2.1.2 Distributing Spam

All interaction with Twitter occurs through authenticated accounts. We distinguish between accounts that are *fraudulent* and explicitly created for spamming versus *compromised* accounts, which belong to legitimate users whose credentials have been stolen. Account compromise can occur due to password guessing, phishing, or mistakenly granting a malicious application privileged access to an account (via OAuth [24] credentials or revealing an account's password). A final approach to distributing spam on Twitter relies on legitimate accounts that post untrusted syndicated content. As we will discuss in Section 5, a number of services have appeared that pay users to post arbitrary content to their profile.

2.2 Detecting Social Network Spam

The diverse array of social network spam and its evasive nature makes it difficult to obtain a comprehensive source of ground truth for measurement. Previous approaches include using blacklists to identify URLs on both Facebook and Twitter directing to spam content [5, 6], deploying passive social networking accounts to act as spam traps [12, 19], and manually identifying spam tweets in trending topics [2]. Each of these approaches introduce a unique bias and error in the type of spam identified. For instance, blacklists preclude URLs that were not reported by users or that failed to appear in email spam traps. Our previous study of Twitter identified that 8% of unique URLs were blacklisted [6], but manual analysis of the same dataset identified 26% of unique URLs directed to spam, indicating blacklists missed a large portion of spam. False positives and negatives also remain a flaw of social spam traps. Stringhini et al. found that passive accounts acting as spam traps received a surprising volume of legitimate traffic, with only 4.5% of friend requests on Facebook originating from spammers, com-

pared to 90% on Twitter [19]. Equally problematic, samples generated from friend requests will omit spam from compromised accounts in addition to spammers who do not form social connections. While manual analysis by experts reduces the potential for error, it is prohibitively expensive for acquiring a large data sample. Our approach of using Twitter's detection algorithm is not without its own bias, which we show in Section 3. As such, we remain cautious of drawing conclusions for *all* spam on Twitter.

2.3 Twitter's Detection Algorithm

Our dataset hinges on Twitter's algorithm for suspending accounts that participate in multiple forms of abusive behavior. While the exact implementation of Twitter's suspension algorithm is not public, many of the parameters are documented in their guidelines and best practices [21]. Accounts will be suspended for frequent requests to befriend users in a short period, reposting duplicate content across multiple accounts, sending unsolicited mentions, posting only URLs, and posting irrelevant or misleading content to trending topics. Other behaviors not clearly related to spam that will result in account suspension include copyright infringement, harassment, and inappropriate content. Given the multiple reasons beyond spam as a grounds for suspension, we validate our dataset in Section 3.1 to confirm the vast majority of suspensions are rooted in spamming behaviors.

2.4 Spam Marketplaces

Within recent years a great deal of effort has been spent on studying the activities of underground economies, and in particular, the spam marketplace. Previous research has examined the hosting infrastructure of scams [1, 7], the organization of email spam campaigns [11], and the economic incentives of spam and malware [9, 18]. Most recently, Levchenko et al. performed an analysis of the infrastructure employed by criminals to monetize email spam, starting from the delivery of an email message and ending with order fulfillment and payment processing [13]. They found that as spammers have become more sophisticated, specialized roles have appeared that separate the marketing of spam products from their actual sale. Organizations with access to pharmaceuticals and replica goods generate affiliate programs, outsourcing the distribution of spam emails to affiliates. As we show in Section 5, a similar diversification of activities is beginning to appear on Twitter, a sign that the Twitter spam market place is maturing.

3. METHODOLOGY

In order to characterize the tools and services that Twitter spammers rely on, we aggregate a dataset of nearly 1.8 billion tweets sent by 32.9 million Twitter accounts over a 7 month period. Of these, we identify 1.1 million accounts suspended by Twitter for abusive behavior. Combined, these accounts sent over 80 million tweets containing 37 million distinct URLs. We manually verify a sample of suspended accounts and find the vast majority were suspended for spamming, providing us with a rich source of ground truth for measuring spam. In addition to our Twitter dataset, we resolve the first redirect of 15 million URLs to deobfuscate a layer of shortening. Finally, for 10 million URLs shortened by *bit.ly*, we download multiple statistics provided by *bit.ly* including clickthrough and, when available, the *bit.ly* account that shortened the URL. A summary of our dataset can be found in Table 1.

3.1 Twitter Dataset

Our Twitter dataset consists of over 1.8 billion tweets collected from Twitter's streaming API [22] during a seven month period from August 17, 2010 to March 4, 2011. We access Twitter's API

Data Source	Sample Size
Tweets	1,795,184,477
Accounts	32,852,752
Distinct URLs	1,073,215,755
Tweets from Suspended Accounts	80,054,991
Suspended Accounts	1,111,776
Distinct URLs from Suspended Accounts	37,652,300
Resolved URLs	15,189,365
Bit.ly URLs	10,092,013
Bit.ly Accounts	23,317

Table 1: Summary of data collected from Twitter, Bit.ly, and from resolving the first redirect of URLs

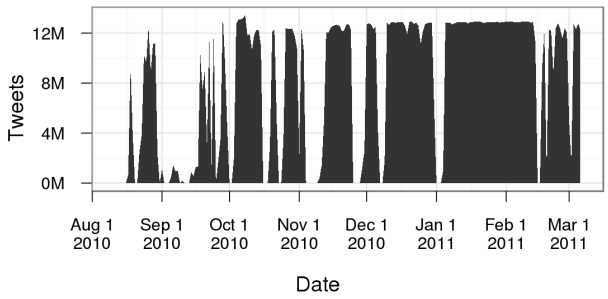


Figure 1: Tweets containing URLs received per day. On average, we receive 12 million tweets per day, with a ceiling imposed by Twitter.

through a privileged account, granting both increased API requests per hour and a larger sample than would be conferred to a default account. We rely on the `statuses/filter` method to collect a sample of public tweets conditioned to contain URLs. For each tweet, we have the associated text of the tweet, the API client used to post the tweet (e.g. web, third-party client), as well as statistics tied to the account who posted the tweet including the account's number of friends, followers, and previous posts. On average, we receive 12 million tweets per day, with a ceiling imposed by Twitter capping our collection at 150 tweets per second. We lack data for some days due to network outages, updates to Twitter's API, and instability of our collection infrastructure. A summary of tweets collected each day and outage periods is shown in Figure 1.

In order to label spam within our dataset, we first identify accounts suspended by Twitter for abusive behavior. This includes spam, aggressive friending, and other non-spam related offenses, as discussed in Section 2.3. Upon suspension, all of an accounts tweets and profile data become restricted and relationships disappear from the social graph. While this provides a clear signal to identify suspended accounts, it also eliminates any possibility of simply downloading an account's history upon suspension. Nevertheless, we are able to reconstruct a composite of a suspended account's activities from all of the tweets in our sample set.

Due to delays in Twitter's account suspension algorithm, we wait two weeks from the last day of data collection before we determine which accounts were suspended by Twitter. This process consists of a bulk query to Twitter's API to identify accounts that no longer have records, either due to deletion or suspension, followed by a request to access each missing account's Twitter profile via the web to identify requests that redirect to *http://twitter.com/suspended*. Of 32.9 million accounts appearing in our sample, 1.1 million were

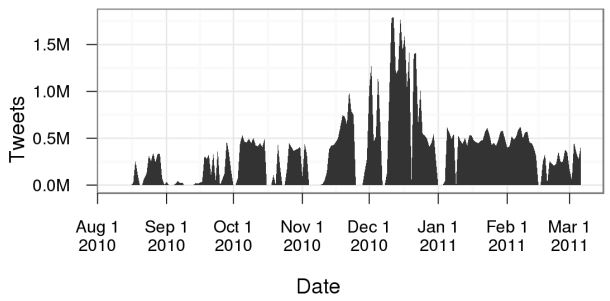


Figure 2: Daily tweet activity of suspended users. Peak activity preceded the holiday season in December.

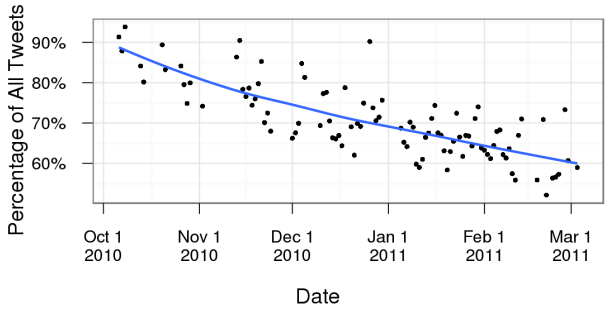


Figure 3: Estimated percentage of all tweets containing URLs we receive per day. Due to Twitter's cap of 12 million tweets per day, we receive a smaller sample size as Twitter grows.

subsequently suspended by Twitter, roughly 3.3% of accounts. These accounts posted over 80 million tweets, with their daily activity shown in Figure 2. We sample a portion of these accounts and manually verify that the vast majority are suspended for spam behavior. Furthermore, we provide an estimate for what fraction of each spam account's tweets appear in our sample, as well as provide an estimate for how many spam accounts go uncaught by Twitter and are thus unlabeled in our data set.

Validating Suspended Accounts are Spammers: When we identify a suspended account, we retroactively label all of the account's tweets in our sample as spam. In doing so, we make an assumption that suspended accounts are predominantly controlled by spammers and are not valid accounts performing unrelated abusive behaviors. To validate this assumption, we draw a random sample of 100 suspended accounts and aggregate every tweet posted by the account appearing in our dataset. We then analyze the content of each tweet to identify common spam keywords, frequent duplicate tweets, and tweet content that appears across multiple accounts. Additionally, we examine the landing page of each tweet's URL, if the URL is still accessible, and the overall posting behavior of each account to identify automation.

Of the 100 accounts, 93 were suspended for posting scams and unsolicited product advertisements; 3 accounts were suspended for exclusively retweeting content from major news accounts, and the remaining 4 accounts were suspended for aggressive marketing and duplicate posts. None of the accounts appeared to be legitimate users who were wrongfully suspended. Presumably, any such false positives would later be resolved by the user requesting their account be unsuspended. From these results, we can discern that the majority of accounts we examine are *fraudulent* accounts created by spammers, though the URLs posted by some of these accounts may direct to legitimate content. We provide further evidence that the accounts in our dataset are created explicitly for spamming rather than compromised or legitimate when we examine the relationships and duration of suspended accounts in Section 4.1. From here on out, we refer to suspended accounts as spam accounts interchangeably.

Validating Active Accounts are Non-spammers: False negatives from Twitter's detection algorithm result in omitting a portion of spam accounts from our analysis. To measure what fraction of spam accounts are missed by Twitter, we randomly sample 200 active accounts and evaluate each account's tweet history using the same criteria we applied to validate spam accounts. Of the 200 accounts, 12 were clearly spammers, from which we can estimate that 6% of active accounts are in fact spammers, with an error bound of $\pm 3.3\%$ at 95% confidence. Consequently, many of our measurements may underestimate the total volume of spam on Twitter and the number of accounts colluding in spam campaigns. For the accounts we manually identified as overlooked spammers, we found no significant distinction between their behavior and that of suspended accounts, leading us to believe they fall below some classification or heuristic threshold that bounds false positives.

Estimating the Likelihood Spammers are Caught: Using our estimates of the number of false positives and false negatives that result from Twitter's spam detection algorithm, we can approximate the algorithm's sensitivity, or the likelihood that a spam account posting URLs will be caught. Of the 31 million accounts that were not suspended, 6% are false negatives, amounting to roughly 1.9 million spam accounts that are overlooked. Another 1 million spam accounts were correctly identified by Twitter's algorithm. Applying the metric for sensitivity:

$$sensitivity = \frac{true\ positives}{true\ positives + false\ negative}$$

we find that only 37% of spam accounts posting URLs on Twitter are caught by the suspension algorithm during the period of our measurement. We note that this estimate is sensitive to the error bound of the false negative rate. Despite the potential for omitting a class of spam from our analysis, we show in Section 3.2 that alternative approaches such as using blacklists to identify spam accounts net an entirely different class of spammers. As such, while our sample may be biased, our analysis provides insights into a large population of spam accounts that have previously been uncharacterized.

Sample Rate: As a final validation step, we measure the fraction of URLs posted to Twitter that we receive in our sample. Our daily sample remains roughly constant at 12 million tweets even though Twitter reports exponential growth [23]. After October 12, 2010, Twitter began to impose a rate limit of 150 tweets per second regardless the actual rate they receive tweets. To measure how this impacts our sample, we take a random sample of 1,600 non-suspended accounts that appear in our dataset and download the entirety of their account history. Of these accounts, 1,245 were still publicly accessible, providing a sample of 798,762 tweets. We then filter out tweets that appear during an outage in our collection or that do not contain URLs, leaving a sample of 32,142 tweets, with roughly 465 samples per day. The daily fraction of these tweets that appear in our sample can be seen in Figure 3 along with a fit curve. At our peak collection, we received 90% tweets containing URLs posted to Twitter. The sample rate has since decreased to nearly 60%.

3.2 Spam URL Dataset

From the 80 million tweets we label as spam, we extract 37.7 million distinct URLs pointing to 155,008 full domains (e.g. *an.example.com*) and 121,171 registered domains (e.g. *example.com*). Given the multitude of shorteners that obfuscate landing pages and no public listing, we attempt to fetch each URL and evaluate whether the HTTP response includes a server-side redirect to a new URL. We only resolve the first such redirect, making no attempt at subsequent requests. In total, we are able to resolve 15.2 million URLs, the remainder of which were shortened by services that have since been deactivated, or were inaccessible due to rate limiting performed by the shortening service.

Blacklist Overlap: In prior work, we examined millions of URLs posted to Twitter that appeared in blacklists [6]. To determine whether the spam we identify from suspended accounts differs significantly from spam detected by blacklists, we examine the overlap of our spam URL dataset with blacklists. We take a sample of 100,000 URLs appearing in tweets regardless of whether they are shortened and a second sample of 100,000 unshortened URLs. We consult three blacklist families: SURBL and all its affiliated lists (e.g. SpamCop, Joewein); Google Safebrowsing, both malware and phishing; and URIBL. If a URL was flagged at any any point in the history of these blacklists from August, 2010 till May, 2011, we consider the URL to be blacklisted. We find only 8% of spam tweet URLs appeared in blacklists and only 5% of unshortened URLs. As such, we believe we present an entirely unexplored subset of spam on social networks from both the perspective of fraudulent accounts as well as non-blacklisted spam.

3.3 Bit.ly URL and Account Dataset

Of the URLs associated with spam tweets in our dataset, over 10 million direct to *bit.ly*, a popular shortening service, or one of its multiple affiliated services (e.g. *j.mp*, *amzn.to*). From *bit.ly*'s public API [22] we are able to download clickthrough statistics for a subset of these URLs found in prominent spam campaigns, and when available, the registered *bit.ly* account that shortened the URL. Roughly 47% of *bit.ly* URLs in our dataset had an associated *bit.ly* account, of which 23,317 were unique.

4. TOOLS OF THE TRADE

Within the amalgam of spam activities on Twitter, we identify a diverse set of tools and strategies that build upon access to hundreds of fraudulent accounts, an array of spam URLs and domains, and automation tools for interacting with Twitter. We explore each of these areas in depth and present challenges facing both spammers and Twitter in the arms race of social network spam.

4.1 Accounts

At the heart of the Twitter spam craft are thousands of fraudulent accounts created for the explicit purposes of soliciting products. 77% of these accounts are banned within a day of their first post, and 89% acquire less than 10 followers at the height of their existence. Yet, within Twitter's hostile suspension environment, spammers are still capable of reaching millions of users through the use of unsolicited mentions and trending topics. We examine a range of properties surrounding spam accounts, including the length of their activity, the rate they send tweets, the social relationships they form, and the stockpiling of accounts.

4.1.1 Active Duration

Spam accounts are regularly suspended by Twitter, but it takes time for Twitter to build up a history of mis-activity before taking action. We measure this window of activity for a sample of 100,000 spam accounts created after our measurement began. We omit accounts that were created during one of our outage periods to reduce bias, though an account appearing at the cusp of an outage period will have its activity window underestimated. For each of these accounts, we calculate the difference between the timestamp of an account's first tweet and last tweet within our dataset, after which we assume the account was immediately suspended. Figure 4 shows a CDF of account activity. 77% of accounts were suspended within a day of their first tweet, with 92% of accounts surviving only three days. The longest lasting account was active for 178 days before finally being suspended. While a minority of accounts are able to persist, we show that rapid suspension impacts both the volume of tweets spammers can disseminate and the relationships they can form.

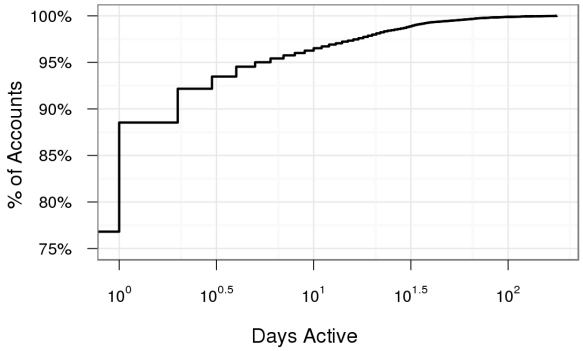


Figure 4: Duration of account activity. 77% of accounts are suspended within a day of their first tweet and 92% within three days.

4.1.2 Tweet Rates

Given the threat of account suspension, we examine whether the rate that spammers send tweets impacts when they are suspended. In order to calculate the total number of tweets sent by an account, we rely on a statistical summary embedded by Twitter in each tweet that includes the *total* number posts made by an account (independent of our sampling). Using a sample of 100,000 accounts, we calculate the maximum tweet count embedded for each account by Twitter and compare it against the account's active duration.

The results of our calculation are shown in Figure 5 along with a fit curve. We identify three clusters in the figure, outlined in ovals, that represent two distinct spamming strategies. The first strategy **(I)** relies on short-lived accounts that flood as many tweets as possible prior to being suspended, representing 34% of our sample. These accounts last a median of 3 days and send 98 tweets. In contrast, a second strategy **(II)** relies on longer lasting accounts that tweet at a modest rate, representing 10% of our sample. While these accounts last a median of 7 days, in the end, they send a nearly equal volume of tweets; a median of 97 tweets per account. The final cluster **(III)** consists of 56% of accounts that are suspended within a median of 1 day and send 5 tweets on average. The reason behind these accounts' suspension is unclear, but it is likely tied to rules beyond tweet count, such as sending URLs duplicated from previously suspended accounts or sharing an email address or IP address with other suspended accounts. While an individual account sending 100 tweets will not reach a large audience, we show in Section 6 that actual spam campaigns coordinate thousands of accounts yielding hundreds of thousands of tweets.

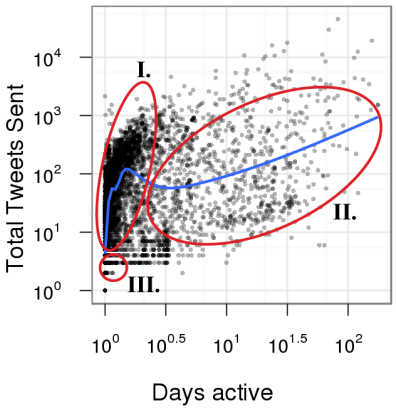


Figure 5: Active duration vs. tweets sent for spam accounts. Two strategies appear: **(I)** burst accounts and **(II)** long-lived, low-daily volume accounts

4.1.3 *Relationships*

The social graph is the focus of regular user interaction on Twitter, yet we find that most spammers fail to form social connections and instead leverage other social networking features to reach an audience. Due to the disappearance of spam accounts from the social graph upon suspension, we are unable to retroactively perform a detailed analysis of the accounts spammers befriended (or whether spammers befriend one another). Nevertheless, each tweet in our dataset includes a snapshot of the number of friends and followers an account held at the time of the tweet. For clarity, we define a *friend* as a second user that an account receives content from, while a *follower* is a second user that receives an account's content. With respect to distributing spam URLs, only followers are important, though spammers will acquire friends in the hope that the relationship will be reciprocated.

To compare relationships formed by both spam and non-spam accounts, we aggregate friend and follower data points for a sample of 100,000 active and suspended users. Figure 6 shows a CDF of the maximum number of followers a spam account acquires prior to suspension. Surprisingly, 40% of spam accounts acquire no followers, while 89% of accounts have fewer than 10 followers. We believe this is due both to the difficulty of forming relationships with legitimate users, as well as a result of the hostile environment imposed by Twitter, where the effort and time required to acquire followers is outpaced by the rate of suspension.

With no followers, spam accounts are unable to distribute their content along social connections. Instead, we find that 52% of accounts with fewer than 10 followers send unsolicited mentions, whereby a personally tailored message is sent to an unsuspecting account that shares no relation with the spammer. Another 17% of accounts rely on embedding hashtags in their spam tweets, allowing spam content to appear in the stream of popular Twitter discussions and through search queries. We examine the success of each of these approaches in Section 6 for a subset of spam campaigns.

For those spam accounts that do form social relationships, their relationships are heavily skewed towards friends rather than followers, indicating a lack of reciprocated relationships. Figure 7 shows the number of friends and followers for spam accounts as well as active accounts presumed to be non-spammers. An identity line in both plots marks equal friends and followers, while a trend line marks the areas of highest density in the scatter plot. Relationships of non-spam accounts center around the identity, while spam

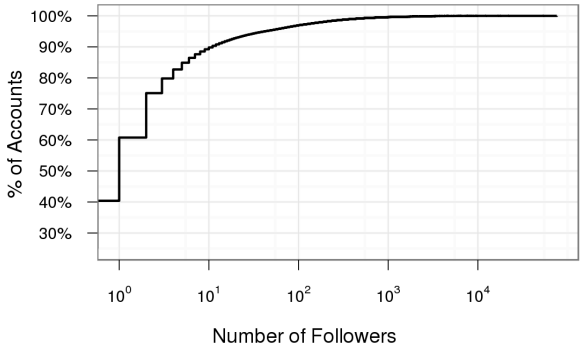
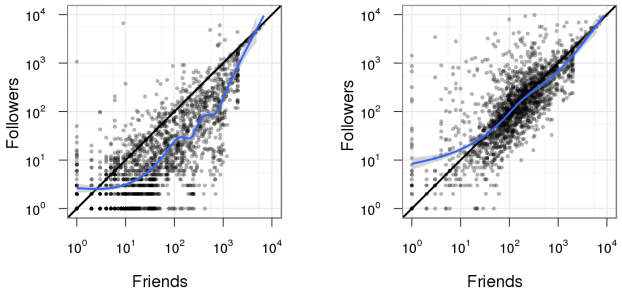


Figure 6: Users following spam accounts. 89% of accounts have fewer than 10 followers; 40% have no followers.



(a) Spam Accounts **(b)** Non-spam Accounts

Figure 7: Friends vs. followers for spam and non-spam accounts. Spammers are skewed towards forming relationships that are never reciprocated.

accounts are shifted right of the identity due to the lack of reciprocated relationships. The modality in both graphs at 2,000 friends results from Twitter imposing a limit on the number of friends possible, after which an account must have more followers than friends to grow their social graph. While 11% of spam accounts attempt to befriend users, either for the purpose of acquiring followers or for obtaining the privilege to direct messages, it is clear that legitimate Twitter users rarely respond in kind.

4.1.4 *Dormancy*

Long dormancy periods where a spam account is registered but never used until a later date hint at the possibility of stockpiling accounts. To measure account dormancy, we select a sample of 100,000 accounts created during one our active collection periods and measure the difference between the account's creation date (reported in each tweet) versus the account's first post in our sample. The results in Figure 8 show that, unsurprisingly, 56% of accounts are activated within a day of their registration. This indicates most spammers create accounts and immediately add them to the pool under their control. However, 12% of spam accounts remain inactive for over a week and 5% for over one month. We highlight this phenomenon further in Section 6 when we present a number of campaigns that stockpile accounts and activate them simultaneously to generate hundreds of thousands of tweets.

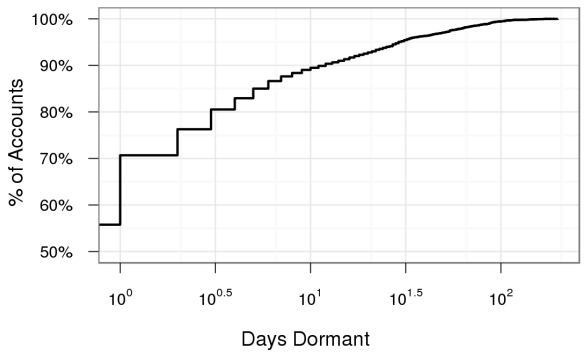


Figure 8: Dormancy duration of accounts. 56% of accounts begin tweeting within the same day the account is created, while 12% lay dormant for over one week, allowing for account stockpiling.

Service Name	% of Tweets	Likelihood Ratio
bit.ly	34.86%	1.41
tinyurl.com	6.88%	2.61
is.gd	2.45%	3.01
goo.gl	2.45%	1.14
ow.ly	2.32%	1.40
dlvr.it	1.99%	1.66
tiny.cc	1.38%	12.36
tiny.ly	1.34%	5.23
3.ly	1.14%	65.55
dld.bz	1.10%	3.71

Table 2: Top 10 public shortening services abused by spammers. Likelihood ratio indicates the likelihood a spammer will use the service over a regular user.

4.2 URLs and Domains

Beyond the necessity of multiple accounts to interact with social networks, spammers also require a diverse set of URLs to advertise. We find that individual spam accounts readily post thousands of unique URLs and domains, simultaneously abusing URL shorteners, free domain registrars, and free web hosting to support their endeavors. Of the 37.7 million spam URLs in our dataset, 89.4% were tweeted *once*. These unique URLs account for 40.5% of spam tweets, while the remaining 10.6% of URLs are massively popular and account for the 59.5% of spam tweets. To understand how spammers are generating URLs, we examine a breadth of properties from the abuse of free services, the diversity of domains, and the overlap of spam URLs with those posted by non-suspended Twitter accounts.

4.2.1 Abusing Shorteners

We find that URL shortening services, such as *bit.ly*, are frequently abused by spammers despite their use of blacklists and spam detection algorithms [3]. In general, URL shorteners simplify the process of generating a variety of unique URLs without incurring a cost to the spammer. URL shorteners also obfuscate the destination of a URL that might otherwise look suspicious to visitors and decrease clickthrough.

Given that any domain can operate a shortening service, we develop a heuristic to identify shorteners used by spammers. Using the first-hop resolution data for 15 million URLs, we identify domains that respond with a server-side redirect (HTTP status code 30x). If a single domain redirects to at least five distinct registered domains, that domain is considered to be a shortening service. Using this criteria, we identify *317 services* that are used in *60% of spam tweets*.

The most popular shorteners abused by spammers are shown in Table 2; 35% of spam tweets are shortened by *bit.ly*, followed in popularity by *tinyurl.com* and a variety of other shorteners with low spam volumes that make up a long tail. For each shortener we compute the bias spammers have towards using the service compared to regular users. First, we calculate $p_1 = p(shortener|spam)$, the probability a spam tweet uses the shortener, and $p_2 = p(shortener \mid nonspam)$, the probability a non-spam tweet uses the shortener. We then compute the likelihood ratio p_1/p_2. This result is strictly a lower bound as our non-spam dataset contains uncaught spam.

As Table 2 shows, all of the top ten shortening services are preferred by spammers, with *3.ly* over 65 times more likely to be used by spammers. The likelihood ratio of a shortener does *not* indicate that more spam URLs are shortened than non-spam URLs. Instead, a likelihood ratio greater than one simply indicates that given the choice of domains available to both spammers and regular Twitter users, spammers are more likely to choose shorteners. Even if popular URL shortening services deployed stronger spam filtering, the presence of hundreds of alternative shorteners and the ease with which they are created makes it simple for spammers to obfuscate their URLs.

4.2.2 Domain Names

Spammers who host their own content require access to hundreds of domains in order to counteract attrition resulting from takedown and blacklisting. Where traditional domain registration carries a cost, we find that Twitter spammers cleverly obtain free hosting and subdomains through the abuse of public services.

Figure 9 visualizes the number of subdomains and the number of registered domains tweeted by each of the 1.1 million spam accounts in our dataset, along with a fit curve showing the densest regions. If an account exclusively posts unique registered domains (e.g. *greatpills.com*), it will appear along the identity line, while accounts that rely on multiple subdomains (e.g. *greatpills.com, my.greatpills.com*) will appear below the identity line. Three distinct approaches to spamming are apparent, outlined in ovals.

The first approach **(I)** consists of 0.13% of spam accounts that abuse free subdomain registration services and blog hosting. These accounts post over 10 subdomains tied to fewer than 10 registered domains, with 0.04% of spam accounts tweeting over 250 subdomains. A second spamming strategy **(II)** consists of using multiple unique domains; we find 1.4% of users tweet over 10 unique registered domains with no additional subdomains, represented by the points scattered along the identify line. The remaining 98.56% of accounts, labeled as **(III)**, tweet fewer than 10 domains in their entire lifetime, appearing as a dense cluster near the origin. Of these clusters, we explore the abuse of free domain registrars and hosting services.

Subdomains:: We identify a number of spammers that rely on free subdomains to avoid registration costs. Services including *co.cc*, *co.tv*, *uni.cc*, and *dot.tk* all allow anyone to register a subdomain that directs to an arbitrary IP address. In total, we find over 350,000 spam URLs directing to these services. The majority of these URLs belong to accounts shown in Figure 9 that post over 250 subdomains.

One particular group of 376 spam accounts advertised over 1,087 subdomains located at *co.cc* with another 1,409 accounts advertising a smaller subset of the same domains. With no limits on subdo-

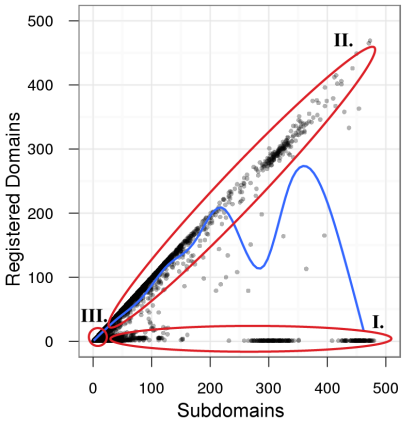


Figure 9: The number of subdomains versus the number of registered domains that URLs posted by a spam account resolve to. Each point corresponds to a single account.

Blog Program	Subdomains	Unique URLs
Blogspot	18,364	249,589
LiveJournal	15,327	54,375
Wordpress	4,290	58,727

Table 3: Top three free blog hosting sites, the number of blogs registered, and the number of unique URLs pointing to the blogs.

main registration services, this *co.cc* campaign displays how spammers can easily circumvent the requirement of domain registration.

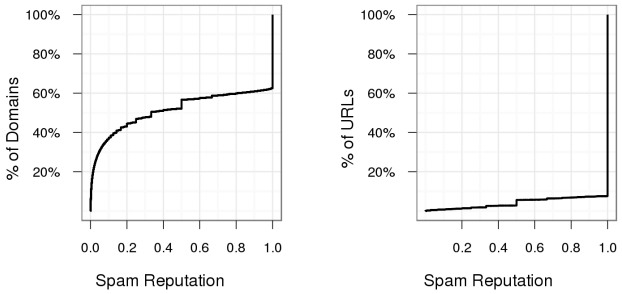
Blog Hosting: Free blog pages from Blogspot, LiveJournal, and Wordpress account for nearly 363,000 URLs; roughly 0.1% of the URLs that appear in our dataset after shortening is resolved. While this may seem minute, Blogspot is the third most popular domain for shortened URLs, highlighting the huge variety of domains used by spammers.

To understand how many blog accounts spammers register, we extract the blog subdomain from each URL and calculate the total number of unique account names that appear, shown in Table 3. Over 18,000 accounts were registered on Blogspot and another 15,000 on LiveJournal. As a whole, we identified 7,500 Twitter spam accounts that advertised one of the three blog platforms, indicating a small collection of spammers who abuse both Twitter and blogging services.

4.2.3 URL and Domain Reputation

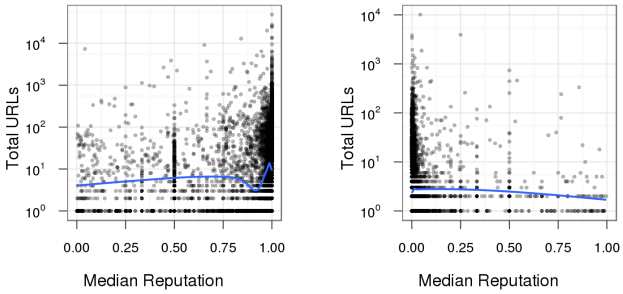
Many of the URLs and domains used by spammers also appear in tweets published by non-suspended users. For instance, of the 121,171 registered domains that appear in spam tweets, 63% also appear in tweets posted by active accounts. To understand whether this overlap is the result of a single retweet of a popular URL or a regular occurrence, we calculate a reputation score for each domain and URL in our dataset. Using all 1.8 billion tweets, we calculate the frequency that a domain or URL appears in spam tweets, and repeat this process for all tweets. The fraction of these two values offers a reputation score in the range $(0, 1)$. A reputation of one indicates a domain or URL appears exclusively in spam, while a reputation near zero indicates the domain or URL was rarely found in spam tweets. We note that due to some spam accounts going unsuspended, our reputation scores underestimate how frequently some domains are used by spammers.

Figure 10 shows the reputation scores for both domains and



(a) Domain Reputation **(b)** URL Reputation

Figure 10: Reputation of spam URLs and domains. 53% of domains appear more frequently in non-spam tweets than spam tweets, though only 2.8% of URLs.



(a) Spam Accounts **(b)** Non-spam Accounts

Figure 11: Comparison of URL reputation and the total spam URLs posted by spam and non-spam accounts, where each point represents a distinct account.

URLs. We find 53% of domains appear more frequently in non-spam tweets than spam, compared to 2.8% of URLs. This indicates that attempting to build a domain blacklist from URLs appearing in spam tweets would be highly ineffective, and more so, attempting to detect unsuspended accounts based on their posting a duplicate spam URLs requires explicit knowledge the URL, not the account posting it, was spam. In fact, 11,573,273 active accounts posted at least one URL also posted by spammers.

To break down this phenomenon further, we examine both the reputation of URLs as well as the frequency that both spam accounts and non-spam accounts post them. Figure 11 shows the median spam reputation of all the URLs posted by an account as well as the number URLs that an account shares in common with spammers. A trend line marks the clusters with the highest density. Spammers are clearly identified by their rarely duplicated, high-spam reputation scores compared to non-spam accounts that post low-spam reputation URLs, though in low frequency. As such, both account behavior as well as the URLs posted would be required to deploy blacklists.

4.3 API Clients

Along with accounts and URLs, spammers require a client to interact with Twitter's API or their web portal. The overall application usage of spam is shown in Table 4. We find 64% of spam

API Name	% of Tweets	Likelihood Ratio
web	58.30%	2.98
twitterfeed	12.39%	1.06
Mobile Web	5.40%	6.07
dlvr.it	2.95%	2.01
hellotxt.com	1.14%	7.89
twittbot.net	1.05%	1.50
EasyBotter	0.98%	4.86
Google	0.83%	0.30
API	0.73%	1.32
www.movatwi.jp	0.72%	1.29
HootSuite	0.71%	0.41

Table 4: Top 10 Twitter clients used by spammers.

originates from the web and mobile web interface, while the remainder of spam is sent from over 10,544 public and custom API clients. We find spammers are nearly three times more likely to use the web interface compared to regular users, and six times more likely to use the mobile web interface. The remaining top 10 applications are automation frameworks that allow scheduled tweets and connecting blog posts to Twitter. Of the API clients we identify, we find over 6,200 are used *exclusively* to send spam, indicating a number of spammers are readily developing custom clients to access their accounts.

5. SPAM-AS-A-SERVICE

We find evidence of an emerging spam-as-a-service market that capitalizes on Twitter, including affiliate programs, ad-based shortening services, and account sellers. Each of these services allow spammers to specialize their efforts, decoupling the process of distributing spam, registering domains and hosting content, and if necessary, product fulfillment. Each of the services we identify reveals a targeted approach to monetizing social networks.

5.1 Affiliate Programs

One aspect of diversification we identify within the underground marketplace is the adoption of affiliate programs, both legitimate and otherwise. From the 15 million URLs we unshortened, we identify two prominent affiliate programs used by spammers: *clickbank.com* and *amazon.com*. Clickbank acts as a middleman, connecting vendors seeking to distribute URLs with affiliates willing to advertise the URLs. Clickbank affiliates are paid based on clickthrough, while vendors are charged a fee. In contrast, Amazon's affiliate program offers up to a 15% commission on purchases made after visitors click on an affiliate's URL. The use of Amazon's affiliate program by spammers blurs the line between what constitutes legitimate advertisement and abuse.

Table 5 shows that over 3.1 million spam tweets directed to Clickbank and nearly 1.2 million to Amazon. While the total number of accounts involved in both affiliate programs is a small fraction of the spam accounts we identify, the abuse of these services hint at an emerging spam-as-a-service market that allows Twitter spammers to exclusively spend their effort on generating accounts and tweets, leaving the task of domain registration and content generation to other parties.

Affiliate programs provide a unique opportunity to explore how individuals are earning a profit by sending spam on Twitter. Assuming each affiliate ID uniquely maps to one spammer, we can group an affiliate's Twitter accounts and the tweets the account's send based on the affiliate ID embedded in each tweet's URL. The results, shown in Table 6, offer a glimpse at the spam infrastructure each affiliate controls. Our analysis reveals a heavily biased environment where a small number of affiliates account for the vast majority of spam. We repeat this same experiment using the 47% of *bit.ly* URLs that contain an associated *bit.ly* account ID. We find that 50% of the 23,317 *bit.ly* accounts control only two or fewer Twitter accounts. Yet, one *bit.ly* account acquired over 5,000 Twitter accounts and sent over 550,000 tweets, revealing again the same biased marketplace that contains thousands of small actors alongside a few major players.

Service	Twitter Accounts	Tweets	Type
Clickbank	16,309	3,128,167	Affiliate
Amazon	8,129	1,173,446	Affiliate
Eca.sh	343	352,882	Shortener
Vur.me	72	9,339	Shortener
Spn.tw	905	87,757	Account
Assetize	815	120,421	Account

Table 5: Programs enabling spam-as-a-service. These include affiliate programs that connect vendors to affiliate advertisers, shorteners that embed ads, as well as account arbitration services that sell access to accounts.

		Tweets		Tw. Accts	
Service	Affiliates	Med	Max	Med	Max
Clickbank	203	565	217,686	2	151
Amazon	919	2	324,613	1	848
Bit.ly	23,317	2	551,200	1	5318

Table 6: Affiliates identified for Clickbank and Amazon along with Twitter accounts they control and the volume of spam they send. Bit.ly accounts reveal a similar result. Both show a biased environment where a small number of spammers account for the vast majority of spam.

5.2 Ad-Based Shorteners

A second form of monetization we identify is the use of syndicated ads from existing ad networks. Ad-based shortening services such as *eca.sh* and *vur.me* provide public URL shortening services, but in return, embed the destination page for shortened URLs in an `IFrame` and display advertisements alongside the original content. Anyone can register an account with the service and will receive a portion of the revenue generated by the additional advertisements. For ad-based URL shorteners, spammers need not control the content they shorten; any major news outlet or popular URL can be shortened, requiring the spammer only handle distribution on Twitter. Within our set of spam, there are over 362,000 tweets sent by 415 accounts using ad-based shorteners, a breakdown of which is provided in Table 5.

5.3 Account Sellers and Arbiters

The final monetization technique we find in our spam dataset are services that sell *control* of accounts as well as sell *access* to accounts. One particular service, called Assetize (since disabled), allowed Twitter users to sell access to their accounts. Assetize drew in over 815 accounts, in turn composing tweets and sending them on each account's behalf. In return, the account's owner would be paid. A similar service called Sponsored Tweets (*http://spn.tw*) is currently in existence and allows anyone to register to have advertisements posted to their account, with 905 such accounts appearing in our spam dataset.

Campaign	Tweets	Accounts	URLs	Hashtags	Mentions	Med. Followers	Med. Tweets
Afraid	14,524,958	124,244	14,528,613	-	11,658,859	2	130
Clickbank	3,128,167	16,309	1,432,680	3,585	542,923	9	108
Yuklumdegga	130,652	2,242	24	11	-	3	83
Amazon	129,602	848	1	-	118,157	22	123
Speedling	118,349	1,947	89,526	4674	870	95	190

Table 7: Summary of major spam campaigns on Twitter. This includes the number of tweets, accounts, unique URLs, unique hashtags, and unique mentions. In addition, we include the median number of followers and tweets for accounts in the campaign.

A second form of spam-as-a-service includes programs that specialize in the sale of Twitter accounts, violating Twitter's Terms of Service [25]. A number sites including *xgcmedia.com* and *backlinksvault.com* purport to register accounts with unique email addresses and create accounts with custom profile images and descriptions. While we cannot directly measure the popularity or impact of these services on Twitter, previous work has examined advertisements for these programs and their associated costs [14]. Both account arbiters and sellers reveal a fledgling market where spammers with content to advertise can obtain access to Twitter accounts without requiring CAPTCHA solvers or other tools to enable automated account creation.

6. SPAM CAMPAIGNS

In this section, we explore five major spam campaigns executed on Twitter that highlight the breadth of tools employed by spammers and the ingenuity of their approaches. Some campaigns are executed by centralized controllers orchestrating thousands of accounts, while others exhibit a decentralized spamming infrastructure enabled by spam-as-a-service programs. Only one of the five campaigns advertises content also found in major email spam campaigns [13], leading us to believe some of the actors in the Twitter spam market are separate from the email marketplace dominated by botnets. A summary of each campaign can be found in Table 7. Due to the multitude of obfuscation techniques used by spammers, there is no simple mechanism to cluster tweets and accounts into campaigns. As such, we describe our methodology for generating campaigns on a per-campaign basis.

6.1 Afraid

The largest campaign in our dataset consists of *over 14 million tweets and 124,000 accounts*. During a period in December when we first identified the campaign, accounts were distributing Amazon affiliate URLs linking to a variety of products. All the URLs distributed by the campaign directed to custom shorteners that have since disappeared, making further analysis impossible. The sheer volume of spam directing to Amazon underscores the blurred line between what constitutes legitimate content compared to traditional email pharmaceuticals and replica goods. As we will show with two other campaigns, spammers are readily capitalizing on the ability to send unsolicited tweets to large audiences on Twitters to push legitimate goods for their own profit.

Despite regular account suspensions, the campaign sustained itself over a 6 month period, relying on unsolicited mentions to reach out to over 11.7 million distinct Twitter users. As Table 7 shows, accounts in the campaign completely ignore the social graph, acquiring a median of two followers throughout their lifetime. Figure 12a shows the creation time, activation time, and suspension time for each of the accounts in the campaign. Most dates for activation and suspension overlap due to the short-lived nature of accounts. Accounts are clearly registered in bulk (as indicated by the vertical lines resulting from duplicate registration dates), sometimes *months* in advance of their final activation, leading us to believe accounts were controlled in a centralized fashion.

Every tweet of the campaign included at least one unique URL along with a random amalgamation of tweet content stolen from other users' tweets, making text-based clustering difficult. We identify tweets that belong to the campaign based on two criteria: a regular expression that captures a textual artifact appearing in the campaign's tweets, and a second expression that captures the re-use of domains across multiple tweets. A more in-depth treatment of our methodology can be found in Appendix A.1. In total, we find over 178 unique domains used exclusively by the campaign, 140 of which rely on *afraid.org* for nameservers. Those domains still being hosted can be resolved, but do not forward traffic to the original campaign landing page.

6.2 Clickbank

Clickbank is one of the highest volume spam-as-a-service programs we identify within our dataset, consisting of over 16,000 Twitter accounts each operating in a decentralized fashion controlled by over 200 affiliates. Nearly 13% of *bit.ly* URLs redirect to Clickbank, making Clickbank the most frequent spam domain directed to by the shortener. Figure 12b shows the prevalence of accounts tweeting Clickbank throughout time. Clickbank URLs appear consistently from the onset of our collection window to its completion, despite accounts being suspended at regular intervals.

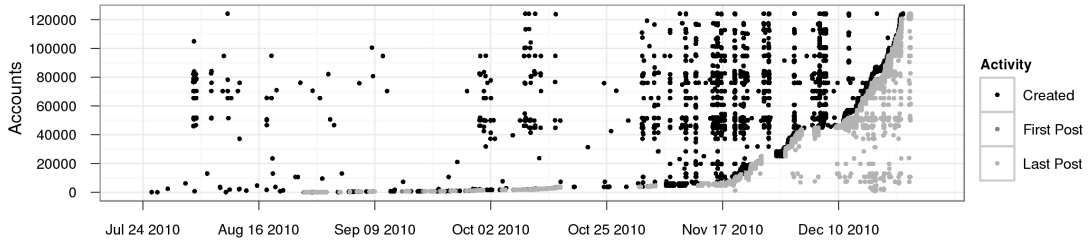
Due to the multiple actors within the campaign, a variety of spamming approaches appear, including the use of unsolicited mentions as well as popular trends. To understand the effectiveness of Clickbank spammers, we take a sample of 20,000 *bit.ly* URLs that direct to the affiliate program and examine their clickthrough. Over 90% of URLs received no clicks at all, though a total of 4,351 clicks were generated for all 20,000 URLs.

We identify tweets belonging to Clickbank participants based on whether the URLs advertised direct to *cbfeed.com* or *clickbank.com*, which serve as intermediate hops for all URLs associated with the service. Our criteria matches both the raw URL appearing in a tweet as well as resolved URLs, provided first-hop data is available.

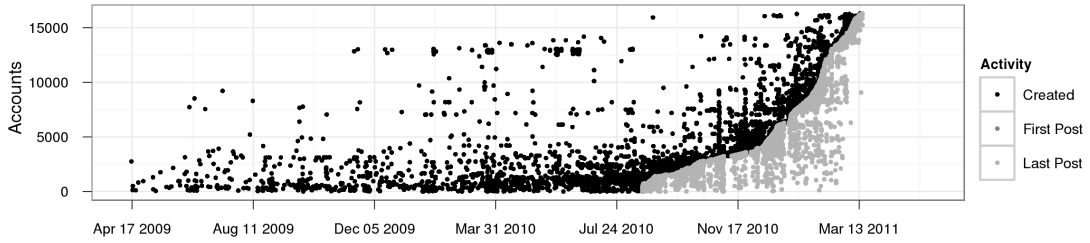
6.3 Yuklemdegga

The Yuklumdegga campaign consists of over 2,200 accounts that pushed pharmacy goods from one of the largest email spam affiliate programs called Eva Pharmacy [10]. Each of the URLs in the campaign resolved to *yuklumdegga.com*, where we identified the HTML store front associated with the Eva Pharmacy program, previously documented by Levchenko et al. [13]. The presence of well-known pharmacy scams on Twitter indicates that some email spam programs are being carried over to Twitter, though we still identify a variety of Twitter spam that does not have a prominent email equivalent.

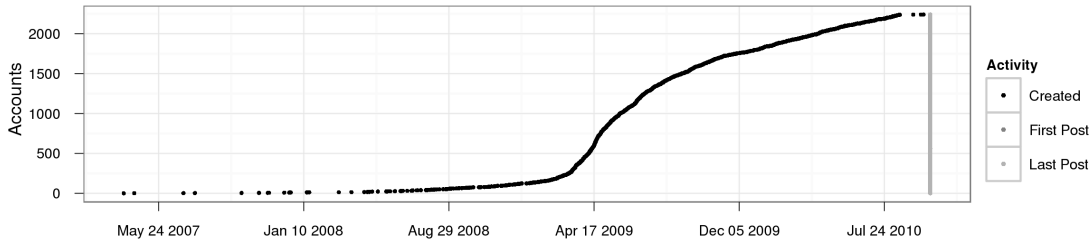
The Yuklumdegga campaign relies exclusively on hijacking trending topics to reach an audience with its tweets, embedding one



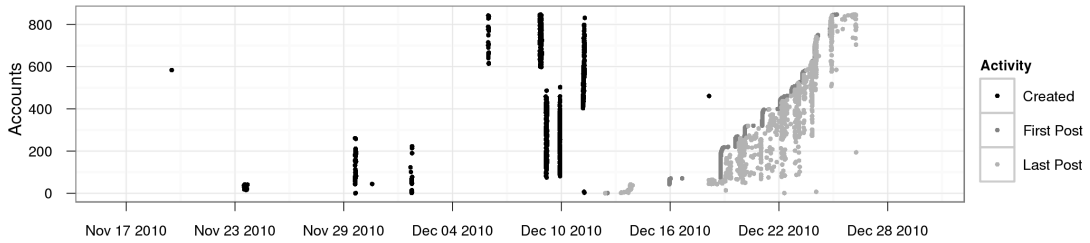
(a) Afraid Campaign



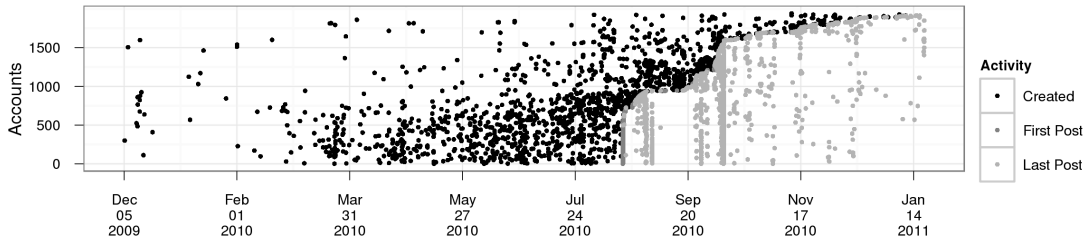
(b) Clickbank Campaign



(c) Yuklumdegga Campaign



(d) Amazon Campaign



(e) Speedling Campaign

Figure 12: Prominent spam campaigns on Twitter

of eleven trends that existed on the single day of the campaign's operation. We find 6 of the campaigns 24 URLs directed to *bit.ly*, with an aggregate clickthrough of 1,982 visitors. Assuming the number of visitors was identical for all 24 URLs of the campaign, we can estimate a total of 8,000 users accessed the pharmacy site, all from reading popular trends.

The preparation of the Yuklumdegga campaign provides a stark comparison to other Twitter campaigns. Figure 12c shows the creation times of the campaigns' accounts, with activation and suspension times overlapped for the single day of the campaign's existence. The bulk of accounts were created *nearly a year* in advance of the campaign's onset. This can result either from the campaign purchasing accounts from a service, or simply creating accounts in a serial fashion until their final activation.

Given that many of the campaign's accounts were created prior to our collection window, we may incorrectly estimate when an account was activated. We determine this is in fact not the case. We calculate the number of posts sent prior to an account's first tweet in our dataset using statistics embedded in each tweet (described in Section 4.1). We find accounts posted a median of 4 posts prior to the campaign's onset, indicating thousands of accounts were stockpiled and then used to send hundreds of tweets.

6.4 Amazon

Like the Afraid campaign, a second Amazon affiliate spam campaign appeared simultaneously during the holiday season. With only 848 accounts, the campaign relied on unsolicited mentions to reach out to over 118,000 Twitter users, each pushing a single URL shortened by *bit.ly*. *Bit.ly* reports the URL received an astounding *107,380 clicks* during the course of the URL's existence. Using our estimate of our sample size for that day (derived in Section 3.1), we can generate an estimate for how many unsolicited mentions result in a visit. Given we received 70% of URLs posted to Twitter in December, roughly 185,000 tweets would have been sent by the campaign. This would indicate over 58% of users clicked on their unsolicited mention, assuming that no other channels advertised the URL and the absence of automated crawlers visiting the URL.

With respect to the accounts participating in the campaign, Figure 12d shows the majority of accounts were created in bulk prior to the holiday and activated in rolling fashion around December 18th. As Twitter banned the active accounts, dormant accounts were then deployed to keep the total accounts active at any point roughly constant. This technique highlights how stockpiled accounts can sustain a campaign through a high-value period.

6.5 Speedling

Speedling (*http://www.speedlings.com/*) is a software program used to generate thousands of blogs that embed advertisements as well as Amazon affiliate URLs generated from a product theme such as cookbooks, games, or any arbitrary keyword. We identify Speedling participants by the API client they use to access Twitter (Speedling, Go Speedling) as well as the format of Speedling's URLs. Further details can be found in Appendix A.2. In total, we find over 89,526 URLs directing to 1,971 domains, all registered to *speedeenames.com*.

As with Clickbank, Speedling represents a decentralized spam campaign consisting of multiple users of the Speedling software. Figure 12e shows the creation and activation time of accounts within the campaign. The vertical lines of final posts indicate mass suspensions on the part of Twitter, yet new accounts are registered, sustaining Speedling's presence on Twitter from the start of our collection period till mid-January.

The monetization approach of Speedling is one of the most interesting compared to the other Twitter campaigns we examine. Visitors are directed to template blog pages listing thousands of Amazon products catering to a specific interest, in addition to a live stream of recommendations from Twitter users that are in fact Twitter accounts controlled by the Speedling operator. Visitors that click on an ad or purchase an Amazon product directly translate into profit, allowing Speedling participants to operate purely through legitimate advertisements. As with the other Amazon affiliate spam we identified, this approach highlights the semi-legitimate nature of some spam campaigns on Twitter, where the distribution approach rather than the landing page or products sold are what distinguish spam from legitimate advertisements.

7. DISCUSSION

In this section, we discuss the implications of our analysis and how our results compare to previous work on social network spam. We also present potential directions for new spam defenses based on our results, describing both in-network solutions as well as addressing the services that facilitate spamming.

Compromised vs. Fraudulent Accounts: Earlier studies of social networks found that 97% of accounts sending spam on Facebook were compromised [5], compared to 84% of accounts on Twitter [6]. In contrast, we find a majority of suspended accounts in our dataset were fraudulent and created for the explicit purpose of spamming. We believe this disparity results from how the datasets for each study were generated. As we showed in Section 3.2, only 8% of the URLs posted by fraudulent accounts appeared in blacklists. These same blacklists served as the source of identifying social network spam in the previous studies, with an apparent bias towards compromised accounts. Our results should thus be viewed in conjunction with these previous studies, offering a wider perspective of the multitude of spamming strategies in social networks.

Blacklists and Spam Traps: With Twitter catching an estimated 37% of spam accounts, a number of studies have examined how to improve this accuracy. These approaches include account heuristics that identify newly created accounts and the lack of social relationships; the identification of unsolicited mentions based on the social graph; and real-time classification of URLs posted to Twitter [2, 12, 17, 19, 20]. Each of these approaches hinges on access to accurate training data, which in practice is often difficult to acquire.

One potential solution for aggregating training data and improving spam detection is to develop Twitter-specific blacklists and spam traps. As previous research has shown, existing blacklists are too slow at identifying threats appearing on social networks, as well as often inaccurate with respect to both false positives and negatives [6, 15, 16]. Even though only 10.6% of URLs in our dataset appear in multiple spam tweets, they account for 59.5% of spam. To capture this re-use, as soon as an account is suspended, the URLs it posted and their associated final landing pages could be added to a blacklist along with the frequency they appeared. If Twitter consulted this blacklist prior to posting a URL, services such as Clickbank would be taken offline, while campaigns that persist despite account suspension would be forced to diversify the URLs they post.

Additionally, rather than suspended accounts outright, Twitter could quarantine tweets from known spam accounts, obtaining access to a steady stream of spam URLs for both classification and blacklisting. While such quarantine is standard practice for email, Twitter has the added difficulty that spammers can easily observe the system to confirm the delivery of their tweets. They can do so by either forming relationships between their accounts to monitor tweet delivery (though this risks Sybil detection [4, 26]), or, alter-

natively, polling the search API confirm whether their spam tweets were indexed. These approaches however incur an additional burden to operating fraudulent accounts.

Beyond Social Networks: The Twitter spam marketplace relies on a multitude of services that include popular URL shorteners, free web hosting, legitimate affiliate programs like Amazon, and illegitimate programs such as Clickbank, Assetiz, and account sellers. While the vast majority of research efforts have targeted spam as it appears on social networks, solutions that disincentivize the abuse of these individual programs would be equally viable. Shortening services, including *bit.ly* and HootSuite, already employ blacklists before URLs are shortened [3, 8]. By monitoring which services underpin the spam ecosystem on Twitter as we do in this study, we can deploy customized countermeasures for each service, reducing the support infrastructure available to spammers.

8. CONCLUSION

This paper presents a unique look at the behaviors of spammers on Twitter by analyzing the tweets sent by suspended users in retrospect. We found that the current marketplace for Twitter spam uses a diverse set of spamming techniques, including a variety of strategies for creating Twitter accounts, generating spam URLs, and distributing spam. We highlighted how these features are woven together to form five of the largest spam campaigns on Twitter accounting for nearly 20% of the spam in our dataset. Furthermore, we found an emerging *spam-as-a-service* market that includes reputable and not-so-reputable affiliate programs, ad-based shorteners, and Twitter account sellers.

In particular, we found that 89% of fraudulent accounts created by spammers forgo participation in the social graph, instead relying on unsolicited mentions and trending topics to attract clicks. Surprisingly, 77% of accounts belonging to spammers were suspended within one day, yet despite this attrition rate, new fraudulent accounts are created to take their place, sustaining Twitter spam throughout the course of our seven month measurement. By examining the accounts controlled by individual spammers as revealed by affiliate programs, we find a handful of actors controlling thousands of Twitter accounts, each pushing a diverse strategy for monetizing Twitter. As a whole, our measurements expose a thriving spam ecosystem on Twitter that is unperturbed by current defenses. Our findings highlight the necessity of better spam controls targeting both abusive accounts as well as the services that support the spam marketplace.

9. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0433702, 0905631, 0842694, 0842695, and 0831501. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work is partially supported by the Office of Naval Research under MURI Grant No. N000140911081.

10. REFERENCES

[1] D. Anderson, C. Fleizach, S. Savage, and G. Voelker. Spamscatter: Characterizing internet scam hosting infrastructure. In *USENIX Security*, 2007.

[2] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting Spammers on Twitter. In *Proceedings of the Conference on Email and Anti-Spam (CEAS)*, 2010.

[3] bit.ly. Spam and Malware Protection. 2009. `http://blog.bit.ly/post/138381844/spam-and-malware-protection`.

[4] G. Danezis and P. Mittal. Sybilinfer: Detecting sybil nodes using social networks. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2009.

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

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

[7] T. Holz, C. Gorecki, F. Freiling, and K. Rieck. Detection and mitigation of fast-flux service networks. In *Proceedings of the 15th Annual Network and Distributed System Security Symposium (NDSS)*, 2008.

[8] HootSuite. Kapow! HootSuite Fights the Evils of Phishing, Malware, and Spam. 2010. `http://blog.hootsuite.com/hootsuite-fights-malware-phishing/`.

[9] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. Voelker, V. Paxson, and S. Savage. Spamalytics: An empirical analysis of spam marketing conversion. In *Proceedings of the 15th ACM Conference on Computer and Communications Security*, 2008.

[10] B. Krebs. Battling the zombie web site armies. `https://krebsonsecurity.com/2011/01/battling-the-zombie-web-site-armies/#more-7522`, 2011.

[11] C. Kreibich, C. Kanich, K. Levchenko, B. Enright, G. Voelker, V. Paxson, and S. Savage. Spamcraft: An inside look at spam campaign orchestration. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2009.

[12] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceeding of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2010.

[13] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Felegyhazi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G. M. Voelker, and S. Savage. Click Trajectories: End-to-End Analysis of the Spam Value Chain. In *Proceedings of the IEEE Symposium on Security and Privacy*, May 2011.

[14] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker. Dirty jobs: The role of freelance labor in web service abuse. In *Proceedings of the 20th USENIX Security Symposium*, 2011.

[15] A. Ramachandran, N. Feamster, and S. Vempala. Filtering spam with behavioral blacklisting. In *Proceedings of the 14th ACM Conference on Computer and Communications Security*, 2007.

[16] S. Sinha, M. Bailey, and F. Jahanian. Shades of grey: On the effectiveness of reputation-based blacklists. In *3rd International Conference on Malicious and Unwanted Software*, 2008.

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

[18] B. Stone-Gross, R. Abman, R. Kemmerer, C. Kruegel,

D. Steigerwald, and G. Vigna. The Underground Economy of Fake Antivirus Software. In *Proceedings of the Workshop on Economics of Information Security (WEIS)*, 2011.

[19] G. Stringhini, C. Kruegel, and G. Vigna. Detecting Spammers on Social Networks. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*, 2010.

[20] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and Evaluation of a Real-time URL Spam Filtering Service. In *Proceedings of the 32nd IEEE Symposium on Security and Privacy*, 2011.

[21] Twitter. The Twitter Rules. `http://support.twitter.com/entries/18311-the-twitter-rules`, 2010.

[22] Twitter. Twitter API wiki. `http://dev.twitter.com/doc`, 2010.

[23] Twitter. Numbers. `http://blog.twitter.com/2011/03/numbers.html`, March 2011.

[24] Twitter. OAuth FAQ. `https://dev.twitter.com/docs/auth/oauth/faq`, 2011.

[25] Twitter. Terms of service, May 2011. `http://twitter.com/tos`.

[26] H. Yu, M. Kaminsky, P. Gibbons, and A. Flaxman. Sybilguard: defending against sybil attacks via social networks. *ACM SIGCOMM Computer Communication Review*, 2006.

APPENDIX

A. IDENTIFYING CAMPAIGNS

This appendix includes a detailed analysis, when applicable, of the clustering criteria used to identify the campaigns discussed in Section 6. The process of finding spam campaigns is a mixture of manual exploration and automated clustering. We begin by independently clustering all spam tweets based on text, URLs and domains, resolved URLs and domains, and the application used to interact with Twitter. After manually identifying the largest and most interesting clusters across the different approaches, we tune our clustering technique on a per-campaign basis to merge related clusters.

A.1 Afraid

Tweets belonging to the Afraid campaign share no textual similarity other than a rare artifact that multiple retweets are included in a single tweet, violating the definition and functionality of a retweet (i.e. exposing a tweet to an account's audience while maintaining attribution). Additionally, many tweets share the same full domain, though domains alone are not enough to capture all tweets belonging to the campaign. We employ a regular expression to identify tweets with numerous embedded retweets and then group them by the domain advertised. Domain clusters with fewer than tens of thousands of tweets are omitted. The subclusters are finally merged, revealing the full scope of the Afraid campaign. A sample of the campaign's tweets are provided below, with emphasis on our labeling criteria.

`@Aguirre_5030 Haha yes for u` **`RT`** `nikivic i love him and i care lol` **`RT`** `dhegracia: I don't love you`
`http://ciqf.t6h.ru``/HENGK`

`@mahi58` **`RT`** `PoiintGod11: Didn't you just tweet about bad english? Lol` **`RT`** `ashLeyGaneshx3: I didnt get no text` **`http://boo.lol.vc`**`/3GbPH`

A.2 Speedling

Due to the decentralized nature of the Speedling product, where anyone can purchase the software program, multiple approaches to spamming appear. As a result, text-based clustering is impossible. Nevertheless, many Speedling participants rely on a Twitter application provided by the software that is uniquely identified through the Twitter API in the source field as **Go Speedling** or **Speedlings** depending the software version. Other participants do not rely on these APIs, but instead use a shortener that only appears in our spam dataset from Speedling participants. Tweets that satisfy any of these criteria are included in the cluster. We note that this provides a strict lower bound on the presence of Speedling spam as some tweets may not match any of these criteria. A more sophisticated approach would be to cluster based on the HTML template of Speedling-generated blogs. However, as we lack HTML due to the retroactive nature of our analysis and link rot, this is impossible in the context of our current study.

Summary Review Documentation for

"Suspended Accounts in Retrospect: An Analysis of Twitter Spam"

Authors: K. Thomas, C. Grier, V. Paxson, D. Song

Reviewer #1

Strengths: Large data set that relies on Twitter identifying malicious accounts instead of simply using blacklists.Well written paper that provides the right amount of details in each area covered and highlights some of the caveats.

Weaknesses: Relies on Twitter to identify suspended accounts. We can infer from the paper that this only covers 35% of the accounts that should have been suspended (sensitivity ratio).

Comments to Authors: Nice paper! In the intro, you mention Amazon and ask yourself what the definition of a spam is. I know where you are coming from. But the content or product promoted by the spam should not influence what should be considered a spam.

About the twitter data set: Are you saying that you only collected tweets that contain URLs? Why limiting yourself to spams that point to URLs. It would be interesting to comment on what you see in spams without URLs.

Giving your characterization of suspended accounts, you could go one step further and see if you would have been able to identify automatically faster and more accounts that should have also been suspended. This is probably future work.

It's great that you think of validating the suspended accounts but also the accounts that are not suspended. But let's do the maths here: 32M accounts, 1M suspended accounts and no false positives, 31M active accounts 6% of the 31M accounts are false negatives = 1.86M.

So sensitivity: TP/ (TP+FN) = 1/2.86 = 35%.I think you need to highlights this in your paper directly. It's a weakness of this paper, but you should not shy away from it. The next question is whether your data set is then significantly biased when you characterize it? It would have been interesting to see some comments about the characterization of the false negatives to see if there are some obvious differences.

You should then better highlight again that caveat and mention your sensitivity ration in your domain reputation section.

Sections 5 and 6 are great additions to the paper.

Reviewer #2

Strengths: A very interesting, insightful analysis of spammer behavior.

Weaknesses: The analysis does not shed much light directly on how the spam can be reduced.

Comments to Authors: I like this paper and I learnt a lot from it about Twitter spam.

My one complaint was that you didn't take the natural next step of telling us, based on your experience, how we (or, Twitter) can deal with this problem more effectively. I am of course assuming that this is a problem that needs to be dealt with (I am not a Twitter user, and so I cannot tell if the spam has reached an annoying level for an average user).

It wasn't clear to me how you actually pulled out the spam campaigns from your tweets. Was it all manual, ad hoc investigation? Did you have any out-of-band information? Or, did you just look at what the affiliates were doing?

I was also surprised to see that you didn't find much overlap between Twitter spam and email spam. Are the players different? Or, maybe I am misinterpreting your statements.

You should do a better job at explaining Figures 5 and 9. Based on the graphs, I do not really see the modes you talk about and had difficulty mapping the text to points in the graph, especially for Figure 9. Perhaps highlight the relevant regions in the graph, instead of the fitted curve line which is not adding any value for me.

While expanding URLs, why did you stop at the first redirect and why didn't you try to counter rate limiting by querying at a slower rate? If scale (too many URLs) was an issue, you could try to expand a subset to see if that improves coverage.

I find it interesting that you found evidence of deactivated shortening services. I've always thought of their use as odd in some contexts, such as in paper bibliographies.

In 2.1, what is an OAuth permission?

Reviewer #3

Strengths: Good datasets, detailed and interesting set of analysis.

Weaknesses: No discussion on how to deal with spam.

Comments to Authors: This is a solid study of spam in Twitter based on a large dataset and detailed analysis. My comments are relatively minor.

The paper characterizes the market place behind using spam in Twitter but ironically does not offer any discussion on how these findings may be used to prevent such an abuse. It seems that a paper deserve a discussion section that elaborates on how to use the findings.

Authors seem to have a few papers related to spam in Twitter, it is useful to clarify how these papers complement each other.

Given the rate of data collection, it appears that authors have used white listed accounts which should be stated.

Reviewer #4

Strengths: The paper did a good job in demonstrating how spamming works on Twitter. Each step in the analysis is well justified and done carefully.

Weaknesses: It is not so clear how the findings in the paper could be used to improve spam detection. Although this point is not sufficient to reject the paper.

Comments to Authors: This is an interesting paper to read. It provides an insight into how spamming works on Twitter. The authors also carefully take into account the false positives/false negatives and sample rates in the dataset. However, as mentioned previously, it is not so clear how these findings can be used to improve the spam detection. The authors might want to provide more discussion relating to this.

There are a few places that are a little hard to follow or could have been improved, listed as follows:

It would be nice if the authors provide a related work section so that the contributions of the papers in relative to the previous work are clearer.

Section 2.1.2 you might want to give a reference for OAuth for readers who may not be familiar with it.

Section 2.3 should refer to Section 3.1, which gives the details of how the confirmation that suspended accounts are mostly spammers is performed.

Section 3.1 Subsection Validating Suspended Accounts are Spammers. When you analyze the content of the tweets, what are the templates mentioned in the "tweets that follow a template"?

Section 4.1.3 How do you determine the accounts that rely on trending topicsto spread spams?

Section 4.2.2 Figure 9 What are the registered domains and full domains?

Section 6.5 Paragraph 3 What does "freeing Speedling participants from fulfilling orders" mean?

Reviewer #5

Strengths: Well written paper that provides a comprehensive description of Spammers and their strategies in Twitter.

Weaknesses: Some assumptions (which the authors admit) on what consists a Spammer and a manual process to verify Spamming activities. These are not significant problems with the paper, since as the authors correctly point out the line of what is spam on twitter is blurred.

Comments to author: I enjoyed reading the paper which I believe nicely describes spammer characteristics and their strategies in Twitter. I particularly liked the analysis of the likelihood ratios and the reputation scores for spam and non-spam accounts. Some minor problems with the paper include:

The phrasing that "manual analysis … of 1.1 million accounts reveals" or "We manually verify that the vast majority of these - 1.1 million accounts –" is a bit weird. You are of course referring to verifying a small sample of those but the phrasing is a bit exaggerated.

As you mention the lines between spam messages are blurred, so a bit more detail on your manual verification process is needed. How do you analyze the content, and what content is deemed spam?

Similarly, when identifying the spam campaigns, it was not clear to me how all the different accounts were grouped together in a campaign? I guess using the particular keywords in each case, but how do you know that this is one and not many campaigns? What if you apply some automated content clustering technique in the spam tweets? Can you identify these or other campaigns?

Finally, I guess an interesting next step would be to design an anti-spam mechanism using the characteristics you discuss in the paper.

Response from the Authors

Reviews of our paper were largely positive, so the changes we made targeted clarifications of how campaigns were identified; the sensitivity of Twitter's suspension algorithm and its impact on our results; and improving explanations on a number of our figures. Additionally, we expanded our discussion of how our work compares to previous studies of social network spam and the separate conclusions we draw.

Overwhelmingly, reviewers requested a discussion on how to move forward on Twitter spam based on our results. Given that there is already a great deal of research on how to improve spam detection in social networks (which we provide an overview of), we instead discuss alternative approaches for combating spam. This includes Twitter-specific blacklisting, spam traps, and removing or preventing the abuse of services that spammers rely upon. Our discussion is informal, simply highlighting a number of potential directions and the impact they would have on reducing spam based on our results and experience studying Twitter.