@spam: The Underground on 140 Characters or Less *

Chris Grier† Kurt Thomas* Vern Paxson† Michael Zhang†

†University of California, Berkeley
{grier, vern, mczhang}@cs.berkeley.edu

*University of Illinois, Champaign-Urbana
kathoma2@illinois.edu

ABSTRACT

In this work we present a characterization of spam on Twitter. We find that 8% of 25 million URLs posted to the site point to phishing, malware, and scams listed on popular blacklists. We analyze the accounts that send spam and find evidence that it originates from previously legitimate accounts that have been compromised and are now being puppeteered by spammers. Using clickthrough data, we analyze spammers' use of features unique to Twitter and the degree that they affect the success of spam. We find that Twitter is a highly successful platform for coercing users to visit spam pages, with a clickthrough rate of 0.13%, compared to much lower rates previously reported for email spam. We group spam URLs into campaigns and identify trends that uniquely distinguish phishing, malware, and spam, to gain an insight into the underlying techniques used to attract users.

Given the absence of spam filtering on Twitter, we examine whether the use of URL blacklists would help to significantly stem the spread of Twitter spam. Our results indicate that blacklists are too slow at identifying new threats, allowing more than 90% of visitors to view a page before it becomes blacklisted. We also find that even if blacklist delays were reduced, the use by spammers of URL shortening services for obfuscation negates the potential gains unless tools that use blacklists develop more sophisticated spam filtering.

Categories and Subject Descriptors

K.4.1 [**Public Policy Issues**]: ABUSE AND CRIME INVOLVING COMPUTERS

General Terms

Security, Measurement

*This material is based upon work partially supported by the NSF under Grants 0433702, CNS-0905631, and CNS-0831535, and by ONR under MURI Grant N000140911081.

CCS'10, October 4–8, 2010, Chicago, Illinois, USA.

1. INTRODUCTION

Within the last few years, Twitter has developed a following of 106 million users that post to the site over one billion times per month [16]. As celebrities such as Oprah, Ashton Kutcher, and Justin Bieber attract throngs of Twitter followers, spammers have been quick to adapt their operations to target Twitter with scams, malware, and phishing attacks [3]. Promising users great diets and more friends, or simply stealing accounts, spam has become a pervasive problem throughout Twitter [8].

Notable attacks on Twitter include the brute force guessing of weak passwords that led to exploitation of compromised accounts to advertise diet pills [26]. Phishing is also a significant concern on Twitter, leading the site to completely redesign the sending of private messages between users to help mitigate attacks [7]. Even though Twitter is vigilant at notifying users and works to stop phishing, spammers continue to create and compromise accounts, sending messages from them to fool users into clicking on scams and harmful links.

Despite an increase in volume of unsolicited messages, Twitter currently lacks a filtering mechanism to prevent spam, with the exception of malware, blocked using Google's Safebrowsing API [4]. Instead, Twitter has developed a loose set of heuristics to quantify spamming activity, such as excessive account creation or requests to befriend other users [22]. Using these methods along with user-generated reports of spamming and abusive behavior, the site suspends offending accounts, withdrawing their presence from the Twittersphere along with all of the account's messages.

In this paper we describe our findings from a large scale effort to characterize spam on Twitter. After collecting a month-long sample of Twitter data, we examine over 400 million public tweets and crawl 25 million unique URLs. Using an assortment of URL blacklists to identify spam, we find over 2 million URLs that direct users to scams, malware, and phishing sites – roughly 8% of all links posted to Twitter. Analyzing the content of spam messages, we provide a breakdown of techniques employed by spammers to exhort Twitter users to click on links. By studying the accounts involved in spamming, we find evidence that spammers primarily abuse *compromised* accounts in their spamming activity, rather than accounts generated solely for the purpose of spamming, which are significantly less prevalent.

Using clickthrough data generated from spam URLs, we examine the success of Twitter spam at enticing over 1.6 million users into visiting spam web pages. We find that the success of spam is directly tied to having a large audience and a variety of accounts to spam from, while use of certain Twitter-specific features also helps increase user traffic. Overall, we find that 0.13% of messages advertised on Twitter will be clicked, *almost two orders of magnitude higher than email spam* [11].

Given the absence of spam filtering on Twitter, we examine whether the use of URL blacklists would help to significantly stem the spread of Twitter spam. By measuring the time period between a blacklist flagging a spam URL and its appearance on Twitter, we find that blacklists in fact lag behind Twitter, with the majority of spam messages appearing 4–20 days before the URLs embedded in the messages become flagged. In contrast, we find over 90% of visits to spam URLs occur within the first two days of posting, indicating that blacklist lag-time is too long to protect a significant number of users against spam. We also examine how spammers can employ URL shortening services to completely evade blacklists, a current problem for Twitter's malware detection.

In summary, the contributions of this paper are:

- We present the first in-depth look at spam on Twitter, based on a detailed analysis of tweets containing over 2 million distinct URLs pointing to blacklisted scams, phishing and malware.
- We analyze the clickthrough rate for spam on Twitter, finding that 0.13% of users exposed to spam URLs click though to the spam web site.
- We identify a diversity of spam campaigns exploiting a range of Twitter features to attract audiences, including large-scale phishing attacks and targeted scams.
- We measure the performance of blacklists as a filter for URLs posted on Twitter, finding that blacklists are currently too slow to stop harmful links from receiving thousands of clicks.
- We develop techniques to identify and analyze two types of spamming accounts on twitter; those created primarily for spamming and accounts compromised by spammers.

We organize the remainder of the paper as follows. Section 2 presents a brief background on spam and an overview of Twitter. Section 3 describes the data we have collected, and Section 4 discusses trends we find in spam tweets, the users who send them, and the clickthrough rate for URLs in tweets. Section 5 discusses techniques for grouping spam into campaigns and examples of successful campaigns. Section 6 presents our evaluation of blacklists, followed by conclusions in Section 7.

2. BACKGROUND

Email spam has an extensive body of research exploring how to identify, characterize, and prevent spam. Common techniques to filter email spam include IP blacklisting [18], domain and URL blacklisting [23, 25, 27], and filtering on email contents [19]. More sophisticated approaches infer the template used by bots to send spam and use the template as a filter [17]. Like many commercial solutions, we use publicly available URL and domain blacklists to identify spam on Twitter and leave the exploration of classification techniques for future work.

Researchers have sought insight into the internal workings of botnets, responsible for much of email spam [10], to measure the success that email spam has at attracting customers. In Spamalytics, the authors are able to infiltrate the Storm botnet and alter the emails being sent, directly measuring the conversion and clickthrough rate of campaigns executed by the Storm botnet [11]. As Twitter is a new medium for spam, we investigate the clickthrough for spam tweets and offer comparison to that of email. We are currently limited to observing clickthrough and cannot determine the final conversion rate for Twitter spam.

The infrastructure used to host spam web sites has also been of interest, where Anderson et al. explore the infrastructure overlap and degree that common hosting arrangements exist for spam campaigns [1]. Wang et al. focus on the redirection chains used by spammers that use search engine optimization techniques to increase traffic [24]. As we will show, redirection services play a role in spam on Twitter and are used for the majority of spam messages sent; however, the recent adoption of URL shortening services on Twitter changes the landscape of interest.

Twitter has recently been the topic of much research, though we are the first to look at the spam and underground behaviors on Twitter. The most relevant work by Kwak et al. examines the structure of social connections on Twitter, as well as the methods trends are propagated [12], but does not examine the thriving spam ecosystem on Twitter. In addition to the studying the social graph, recent work on social network spam uses machine learning to classify spam tweets [13], determine Twitter influence [2], and classify spam MySpace profiles [9].

Where traditional email spam requires access to bulk lists of email addresses, social network spam requires the generation or subversion of user accounts with access to large groups of friends and social circles. Without access to relationships with other users, a message cannot be propagated. The challenge of a successful spam campaign in Twitter is thus two fold: obtaining enough accounts to carry out a campaign before the accounts involved are suspended, and having enough fresh URLs to evade heuristic detection for excessively posting the same link. Before exploring the scope of spam activity in Twitter, we present a brief overview of how Twitter operates and Twitter-specific features spammers have at their disposal.

2.1 Anatomy of a Twitter spammer

A generic profile on Twitter consists of three components: the account's *tweets*, *followers*, and *friends*.

Tweets: A tweet is a colloquialism used by Twitter to describe a status update, analogous to an email's body. Twitter restricts these updates to 140 characters or less, limiting the amount of information spammers can embed in a single tweet as well as the text that can be considered for spam filtering. To facilitate the posting of URLs in tweets, URL shortening services are commonly used and provide redirection services from a short URL of around 20 characters to an arbitrary URL.

Followers: An account's followers are the set of users that will receive a tweet once it is posted, akin to the `To` field of an email. The challenge for spammers is to obtain a large following, allowing the spammer to advertise a single tweet to thousands of users. Users must subscribe as a spammer's follower before receiving tweets; a spammer cannot force his messages to be viewed by other users.

Friends: Relationships in Twitter are not bidirectional, meaning a user can receive tweets from a friend without revealing their own tweets. Friends are the set of users an account subscribes to in order to obtain access to status updates. In the case of spammers, having friends provides no benefit in generating traffic. However, spammers will befriend multiple victims in the hope some will reciprocate the relationship, opening up a channel for communication.

2.2 Twitter features

In addition to account components, there are a number of Twitter-specific features that can be used in tweets to home in on a specific user or reach a wider audience, including *mentions*, *retweets*, and *hashtags*.

Mentions: To address a particular user, *@username* is included in

a tweet, referencing the user directly. For users with public timelines, mentions appear in a user's timeline regardless of if the user is following the sender. This allows users to quickly identify tweets directed at them (though still broadcast to the sender's followers).

Example: *@justinbieber PLEASE FOLLOOWW MEEE!!! <3333*

Retweets: Retweets on Twitter are a form of attribution, where *RT @username* or *via @username* denote that the tweet text originally appeared on another user's profile. Retweets build on the authority of another user and are used to increase the volume of followers who see a tweet.

Example: *RT @JBieberCrewz: RT this if u <3 justin bieber*

Hashtags: In addition to mentioning users, tweets can include tags to arbitrary topics by including a hashtag *#topic*. If enough users pick up on the topic it will appear in the list of *trending topics*, allowing tweets to be syndicated to all of Twitter. As anyone can contribute to a topic, spammers can latch onto currently trending topics, injecting unsolicited messages into the feed.

Example: *Get free followers #FF #Follow Justin Bieber*

2.3 Presenting tweets to users

Each Twitter user is provided with a customized timeline of tweets generated from content posted by friends. When using the Twitter web page to view a friend's message, a single tweet contains the tweet text, the friend's name and icon, the time posted, geo-location data, and the application used to post the tweet. If a link is posted, these attributes are the only information available for user to base their decision on whether to click the link. As simply visiting a website can lead to the installation of malware, this is a potentially dangerous situation.

3. DATA COLLECTION

Understanding spam behavior on Twitter requires a large-scale, real-time framework for detecting and tracking spam accounts. In this section, we describe the development of our Twitter monitoring infrastructure and the use of URL blacklists to identify spam. Our infrastructure focuses on analyzing the techniques employed by spammers to generate click traffic and attract an audience, in addition to tracking the use of obfuscation and redirects to mask potentially suspicious web pages.

Within the broad spectrum of spam, we monitor three different categories: malware, phishing, and scams. A spam URL is classified as malware if the page hosts malicious software or attempts to exploit a user's browser. Phishing pages include any website attempting to solicit a user's account credentials, many of which specifically target Twitter credentials. Lastly, we define a scam as any website advertising pharmaceuticals, software, adult content, and a multitude of other solicitations.

3.1 Twitter monitoring

To measure the pervasiveness of spam, we develop a Twitter monitoring framework that taps into Twitter's Streaming API[1] and collect roughly seven million tweets/day over the course of one month. We collect data from two separate taps, one targets a random sample of Twitter activity while the second specifically targets any tweets containing URLs. The random sample is used to generate statistics about the fraction of URLs in tweets and general Twitter trends, while the URL stream is used for all other measurements.

Once a tweet appears in the URL stream, we isolate the associated URL and use a custom web crawler to follow the URL through HTTP status codes and META tag redirects until reaching the final *landing page* at a rate of roughly ten landing pages per second; currently, JavaScript and Flash are not handled due to the sheer volume of traffic that must be processed and the complexity required to instrument these redirects. While crawling URLs, each redirect is logged, allowing us to analyze the frequency of cross-domain and local redirects, but more importantly, redirect resolution removes any URL obfuscation that masks the domain of the final landing page. We record the number of redirects and the URLs in each sequence.

3.2 Blacklist detection

To automatically identify spam, we use blacklists to flag known spam URLs and domains. We regularly check every landing page's URL in our data set against three blacklists: Google Safebrowsing, URIBL, and Joewein [6, 23, 25]. Each landing page must be rechecked multiple times since blacklists may be slow to update in response to new spam sites. URLs and domains blacklisted by Google indicate the presence of phishing or malware, while URIBL and Joewein specifically target domains present in spam email and are used by anti-spam software to classify email messages. Once a landing page is retroactively marked as spam, we analyze the associated spam tweets and users involved in the spam operation. We have found that URIBL and Joewein include domains that are not exclusively hosting spam; we created a white-list for popular domains that appear on these blacklists and verified that the domains primarily host non-spam content.

3.3 Data summary

Our data collection spans one month of Twitter activity from January to February, 2010. During this time we gathered over 200 million tweets from the stream and crawled 25 million URLs. Over three million tweets were identified as spam. Of the URLs crawled, two million were identified as spam by blacklists, 8% of all unique links. Of these blacklisted URLs, 5% were malware and phishing, while the remaining 95% directed users towards scams. To understand blacklist performance, we manually inspected a random sample of distinct URLs from tweets, finding that 26% of URLs pointed to spam content, with an error margin of 5% at 95% confidence. To manually classify tweets, one of the authors clicks on the URL in a tweet and decides if the URL is spam based on the content of the web page. Compared to the 8% detected by blacklists, a significant proportion of spam URLs are never seen in blacklists, a challenge discussed in greater detail in Section 6. Over 90% of Twitter users have public accounts [15], and we also collect the complete history for over 120,000 users with public accounts, half of which have sent spam identified by our blacklists; the history is an additional 150 million tweets sent by these users.

In the event *bit.ly* or an affiliated service is used to shorten a spam URL, we use the *bit.ly* API[2] to download clickthrough statistics and click stream data which allows us to identify highly successful spam pages and the rate of traffic. Of the spam links recovered, 245,000 had associated clickthrough data, totaling over 1.6 million clicks. Using all of the links recovered during crawling, we present an analysis of the techniques employed by spammers, using clickthrough statistics when available, to measure effectiveness.

4. SPAM ON TWITTER

With over 3 million tweets posted to Twitter directing users to spam detected by popular blacklists, we present an analysis of the categories of spam appearing on Twitter and what techniques are

[1] http://apiwiki.twitter.com/Twitter-API-Documentation

[2] http://code.google.com/p/bitly-api/wiki/ApiDocumentation

Category	Fraction of spam
Free music, games, books, downloads	29.82%
Jewelery, electronics, vehicles	22.22%
Contest, gambling, prizes	15.72%
Finance, loans, realty	13.07%
Increase Twitter following	11.18%
Diet	3.10%
Adult	2.83%
Charity, donation scams	1.65%
Pharmacutical	0.27%
Antivirus	0.14%

Table 1: Breakdown of spam categories for spam on Twitter, based on tweet text.

being employed to reach audiences. To measure the success of Twitter spam, we analyze clickthrough statistics for spam URLs, estimating the likelihood a spam tweet will be clicked by a follower. Finally, as spammers must coerce Twitter members into following spam accounts, we analyze tweeting behavior to differentiate between automated spamming bots and compromised accounts that have been used to send spam, finding the vast majority of spammers appear to be compromised accounts or unwitting participants in spam distribution.

4.1 Spam breakdown

Aggregating all of the spam tweets identified by our system, we generate a list of the most frequent terms. We then manually classify each term into a spam category when a clear distinction is possible, in turn using the terms to classify all of our spam tweets. Roughly 50% of spam was uncategorized due to using random terms; the breakdown of the remaining 50% of tweets is shown in Table 1. While the typical assortment of scams present in email carry over to Twitter, we also identify Twitter-specific advertisements that sell Twitter followers or purport to give an account free followers. This unique category makes up over 11% of categorized Twitter spam, while the remainder of spam is dominated by financial scams, games, sale advertisements, and free downloads.

With only 140 characters for spammers to present a message, we analyze what Twitter-specific features appear in tweets with blacklisted URLs compared to those of regular users. To act as a control, we select two samples of 60,000 tweets, one made up of any tweet appearing in our stream, while the second sample is generated from only tweets containing URLs. Each tweet is parsed for mentions, retweets, and hashtags, the results of which can be seen in Table 2.

The random sample of tweets is dominated by conversations between users, as indicated by 41% of sample tweets containing mentions. Compared to the sample of tweets containing URLs, spam tweets are only slightly less likely to use Twitter features, with the exception of malware and phishing tweets, where hashtags make up 70% of spam. To understand the motivation for spammers to use these features, we present an analysis of how hashtags, retweets, and mentions are being used by spammers.

Call outs: Mentions are used by spammers to personalize messages in an attempt to increase the likelihood a victim follows a spam link. Mentions can also be used to communicate with users that do not follow a spammer. In our data set, 3.5-10% of spam tweets rely on mentions to personalize messages, the least popular feature compared to hashtags and retweets.

Example: *Win an iTouch AND a $150 Apple gift card @victim! http://spam.com*

Source	#	@	RT	#,@	#,RT
Google	70.1%	3.5%	1.8%	0.1%	0.3%
Joewein	5.5%	3.7%	6.5%	0.2%	0.5%
URIBL	18.2%	10.6%	11.4%	1.5%	1.3%
Tweet	13.3%	41.1%	13.6%	1.8%	2.3%
Tweet, URL	22.4%	14.1%	16.9%	1.6%	2.4%

Table 2: Feature frequency by blacklist for mentions (@), retweets (RT), and hashtags (#), compared to a random sample of tweets and a random sample of tweets containing URLs.

Retweets: Of the spam tweets we observe, roughly 1.8-11.4% are retweets of blacklisted URLs. We identify four sources of spam retweets: retweets purchased by spammers from respected Twitter members, spam accounts retweeting other spam, hijacked retweets, and users unwittingly retweeting spam. Of the sources, we are able to differentiate instances of purchased tweets, discussed further in Section 5, and hijacked retweets which we discuss next.

Example: *RT @scammer: check out the Ipads there having a giveaway http://spam.com*

Tweet hijacking: Rather than coercing another account to retweet spam, spammers can hijack tweets posted by other users and retweet them, prepending the tweet with spam URLs. Currently, there are no restrictions on Twitter on who can retweet a message, allowing spammers to take tweets posted by prominent members, modify them, and repost with spam URLs. By hijacking tweets from prominent Twitter users, spammers can exploit user trust in retweets. Analyzing retweets for prepended text, we find hijacking constituted 23% of phishing and malware retweets, compared to 1% of scam retweets.

Example: *http://spam.com RT @barackobama A great battle is ahead of us*

Trend setting: Hashtags are used to simplify searches for content, and if enough users tweet the same hashtag, it becomes a *trending topic*. The anomaly of 70% of phishing and malware spam containing hashtags can be explained by spammers attempting to create a trending topic, generating over 52,000 tweets containing a single tag. Searching for hashtags that exclusively appear in spam tweets, we identify attempts to initiate a trend. Of the total trends we identify, roughly 14% appear to be generated exclusively by spammers.

Example: *Buy more followers! http://spam.com #fwlr*

Trend hijacking: Rather than generating a unique topic, spammers can append currently trending topics to their own spam. Anyone who searches for the topic will then encounter the spam message, interspersed with other non-spam generated by Twitter users. Using this technique, spammers no longer need to obtain followers and instead ride on the success of other topics. Analyzing the list of trending topics from a set of random tweets, we find that roughly 86% of trends used by spammers also appear in benign tweets, with popular trends at the time including #haiti, #iranelection, #glee, and the #olympics.

Example: *Help donate to #haiti relief: http://spam.com*

4.2 Spam Clickthrough

In the event an account spams URLs shortened with *bit.ly*, we can recover clickthrough statistics for the link and analyze the linear correlation of clickthrough with other features such as followers and tweet behavior. Of the blacklisted domains we identify, we observe the clickthrough data for nearly 245,000 URLs. Roughly 97.7% of URLs receive no clicks, but those that do accumulate

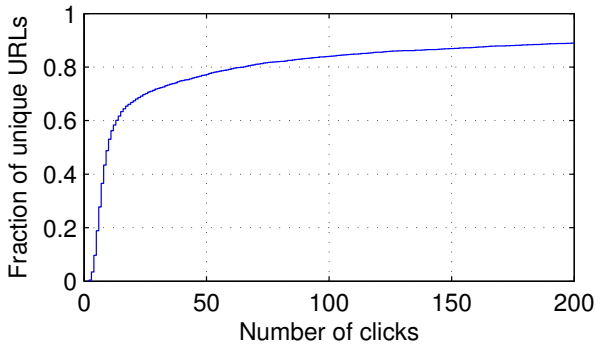


Figure 1: Clickthrough for spam URLs posted to Twitter. Only the 2.3% of URLs that generated any traffic are shown.

over 1.6 million visitors, indicating that spam on Twitter is by no means unsuccessful. Of links that generate any traffic, 50% of the URLs receive fewer than 10 clicks, as shown in Figure 1, while the upper 10% of URLs account for 85% of the 1.6 million clicks we observe. These highly successful URLs are dominated by phishing scams that have pervaded Twitter in recent months [8], and we discuss this further in Section 5.

Using the 2.3% of URLs that receive any traffic, we calculate the linear correlation for clicks and the number of accounts tweeting a link, the aggregate followers that could view the link, and lastly the number of times the link was tweeted, broken down into disjoint combinations of features (*RT, @, #*). Unsurprisingly, the features with the largest coefficient of correlation ($\rho > 0.7$) are the number of accounts involved in spamming and the number of followers that receive a link, both of which directly impact the overall number of potential impressions. In addition to audience volume, we found that the use of hashtags ($\rho = .74$) and retweets with hashtags ($\rho = .55$) is correlated with higher clickthrough rates. In practice, the use of such features is rare, as previously shown in Table 2, but their dominance amongst 70% of phishing and malware tweets bolsters their correlation to successful clickthrough. Surprisingly, the number of times spam is tweeted shows a low coefficient of correlation to clickthrough ($\rho = .28$), indicating that repeatedly posting a link does little to increase traffic.

To understand the effectiveness of tweeting to entice a follower into visiting a spam URL, we measure the ratio of clicks a link receives compared to the number of tweets sent. Given the broadcast nature of tweeting, we measure *reach* as a function of both the total tweets sent t and the followers exposed to each tweet f, where reach equals $t \times f$. In the event multiple accounts with potentially variable number of followers all participate in tweeting a single URL, we measure total reach as the sum of each individual account's reach. Averaging the ratio of clicks to reach for each of the 245,000 URLs in our *bit.ly* data set, we find roughly 0.13% of spam tweets generate a visit, orders of magnitude higher when compared to clickthrough rates of 0.003%–0.006% reported for spam email [11].

There are a number of factors which may degrade the quality of this estimate. First, our data set exclusively targets *bit.ly* URLs which may carry an inherent bias of trust as the most popular URL shortening service [20]. Secondly, click data from *bit.ly* includes the entire history of a link, while our observation of a link's usage only account for one month of Twitter activity. If a link is tweeted prior to our study, or all repeated tweets do not appear in our 10% sample, reach may be underestimated. We attempt to correct for this possibility by measuring the number of times a tweet is repeated using the entire history of 50,000 accounts, finding on average a tweet will appear 1.24 times, with 93% of tweets being unique. This adjustment is factored into the reach of our earlier calculations, but we still caution our estimate of tweet clickthrough as a rough prediction.

Twitter's improved clickthrough rate compared to email has a number of explanations. First, users are faced with only 140 characters in which to base their decision whether a URL is spam. Paired with an implicit trust for accounts users befriend, increased clickthrough potentially results from a mixture of naivety and lack of information. Alternatively, previous estimates of email clickthrough implicitly expect all emails to be viewed. In practice, this may not be the case, resulting in users never being presented the option to click on spam. This same challenge exists in identifying whether a tweet is viewed, but the rates that users view tweets versus emails may differ.

Regardless the underlying cause, Twitter's clickthrough rate makes the social network an attractive target for spammers; with only loose spam filtering in place, spammers are free to solicit throughout the Twittersphere. Furthermore, the computational time of broadcasting tweets is pushed off on Twitter's servers compared to email spam which requires access to large quantities of bots. After a spammer generates a Twitter following, messages can easily be distributed to thousands of followers with a minimal amount of effort.

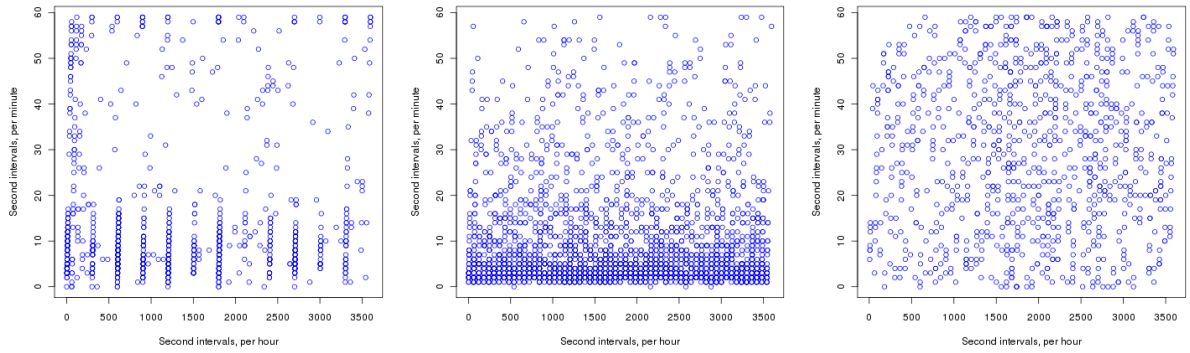
4.3 Spam Accounts

Without Twitter accounts, spammers are incapable of promoting their landing pages. To understand the types of accounts involved in spamming, we define two categories for users flagged as tweeting blacklisted links. The first is the *career* spamming account created with the express purpose of promoting spam. In contrast, a *compromised* account was created by a legitimate user and at some point in time compromised through the use of phishing attacks, malware, or simple password guessing. To differentiate between the two, we develop an array of tests that analyze an account's entire tweet history, finding that the majority of spam on Twitter originates from compromised accounts, not career spammers. It is important to note these tests are not designed to detect spamming accounts and replace blacklists as they can easily be evaded by an adversary. Instead, we rely on these classification techniques solely to help us understand the ecosystem of spam on Twitter.

4.3.1 Career spamming accounts

We develop two tests that indicate if an account is a career spammer, manually verifying the accuracy of each test on a random sample of both spam and likely non-spam accounts. The first test analyzes tweet timing, based on the assumption that legitimate account tweets overall reflect a uniform (Poisson) process. The second test measures the entropy of an account's tweets, identifying users that consistently tweet the same text or link.

χ^2 test on timestamp: Our first test examines tweet timestamps to identify patterns in the minutes and seconds for when a tweet was posted. We represent timestamps for an individual account using vectors corresponding to the seconds value of each hour and seconds value of each minute. We then use a χ^2 test to compute the p-value for these vectors for their consistency with an underlying uniform distribution. For example, a p-value of less than 0.001 indicates less than 0.1% chance that a user posting as a Poisson process generated the sequence. For our evaluation, we treat a p-value of less than 0.001 for either vector as evidence that the user has demonstrably failed the test. Such user tweet patterns very likely reflect automation, leading to postings at regularized times. We



(a) (b) (c)

Figure 2: Scatter plots of times of tweets for three users deemed to not post uniformly. The x-axis gives the minutes value of each hour and y-axis gives seconds. In (a), the user posts at regular intervals – approximately every five minutes. The account in (b) tends to tweet toward the begining of each minute, indicated by the prevalence of points low on the y-axis. For (c), the pattern is less obvious but still caught by the χ^2 test as indicating regularized tweeting with respect to the hour (x-axis).

deem such accounts as likely career spammers. Figure 2 shows examples of the minutes and seconds for three accounts that fail the test. We manually assessed dozens of accounts that both passed and failed this test, including both inspecting the contents of their tweets and their tweeting patterns over time, finding that it is highly accurate in finding what appear to be career spammers.

Tweet text and link entropy: For each spam account, we examine the account's tweets history to identify instances where the text and links posted are dominated by repetition, which we measure by calculating entropy. The test begins by binning the text and URLs posted by an account into distinct bins and calculating the entropy of the resulting distribution for the text and URL. If there is no repetition, then the entropy is equivalent to a uniformly random set of the same size. We then calculate relative entropy as the ratio of observed entropy to the entropy of a uniformly random set of the same size, finding that a relative entropy value less than 0.5 indicates strong repetition. For users that do not repeatedly post the same tweet, relative entropy is close to one.

Using the entire tweet history of a sample of 43,000 spam accounts, each with over 100 tweets per user, we find that roughly 16% of accounts tweeting at least one blacklisted link are career spammers. To gauge the false negative rate of our classification, we manually inspect 99 accounts that passed both the χ^2 and entropy tests to determine a breakdown of the non-career spamming accounts. Of the 99 samples, 35 are not categorized due to tweets appearing in a foreign language and another 5 had been suspended, prohibiting our access to the account's tweet history and reducing our sample size to 59 accounts. Of these, 5 were clearly career spammers that had evaded detection, roughly 8.5% of accounts, with an error bound of 7% at 95% confidence.

To understand why the majority of spam accounts passed both tests, we perform a second test to determine how many blacklisted URLs an average account tweets. For each account in our sample of 43,000, we selected 10% of URLs from the account's history and crawled them to determine the final landing page. Using our blacklists, we identified 304,711 spam landing pages, roughly 26% of URLs crawled. The majority of spam accounts tweeted only 2 spam messages, while the remainder of their tweets appeared to be benign URLs and purely text tweets posted at random intervals. Given the low number of spam URLs, we believe the vast majority of accounts tweeting blacklisted URLs are not career spammers, indicating a potential for compromised accounts.

4.3.2 *Compromised spamming accounts*

With the majority of spamming accounts passing both the χ^2 and entropy tests used to identify automated behavior, we are left with two possibilities for non-career accounts. First, an account could have been compromised by means of phishing, malware, or simple password guessing, currently a major trend in Twitter [26]. As most non-career accounts tweet a limited number of spam URLs, the short lifetime of a compromise can result from Twitter detecting the compromise and notifying the user involved, as occurs with phishing attacks, or the user might identify suspicious activity within their tweet timeline and takes defensive action. Alternatively, given the limited number of spam URLs posted, an account's owner may have tweeted the URLs unintentionally, unaware that they were spam. Given that we expect a non-career spammer to tweet 20 spam URLs, it is unlikely an account mistakenly posts spam so frequently, leading us to believe accounts are in fact compromised.

Compromised accounts present spammers with an attractive means of exploiting trust, using a victim's credibility to push spam out to followers. Furthermore, by taking control of a victim's account, spammers are relieved of the effort of coercing users into following spam accounts. For non-career accounts that tweet malware and phishing URLs, we have strong evidence indicating the accounts involved are likely compromised users. In particular, we identify two major schemes to steal accounts, including phishing pages that purport to provide followers and the Koobface botnet which spreads through URLs in tweets [5]. For accounts identified as tweeting spam domains found in the URIBL and Joewein blacklists, we have less direct evidence indicating accounts were compromised, though there have been examples of such behavior reported [26].

Using a fake account to act as a spam trap, we entered our account information into one of the most frequently spammed phishing sites that was blacklisted by Google's Safebrowsing blacklist. Once phished, the account was used to further advertise the same phishing scam in addition to other spam domains. By searching for these spam URLs in our data set, we identified over 20,000 ac-

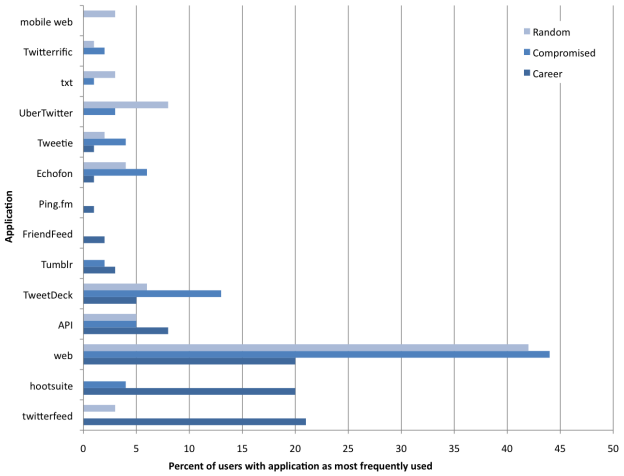


Figure 3: Most frequently used applications per-account for compromised, spamming and a random sample of accounts. Career spammers use different applications than compromised users, which are closer to the random set.

counts that were affected, 86% of which passed our career spammer test.

Further evidence that Twitter accounts are being compromised comes from the spread of Koobface malware which hijacks a victim's Twitter account and tweets on his behalf. During a concerted effort to infiltrate the Koobface botnet, we constructed search queries to find compromised accounts on Twitter and monitored the spread on Twitter during a month of collection. We identified 259 accounts that had tweeted a link leading to a landing page that attempted to install Koobface malware, indicating that these accounts had already been compromised by the botnet and were being used to infect new hosts [21].

These two cases highlight that compromises are occurring on Twitter with the explicit purpose of spreading phishing, malware, and spam. With Twitter credentials being sold in the underground market [14], evidence is mounting that Twitter accounts with large followings are viewed as a commodity, giving access to a trusting audience more likely to click on links, as indicated by our click-through results.

4.3.3 Spam Tools

To understand how spammers are communicating with Twitter, we analyze the most popular applications amongst spam accounts used to post tweets. Using information embedded in each tweet, we aggregate statistics on the most popular applications employed by spammers, comparing these results to a random sample. Figure 3 shows that career spammer application usage is dominated by automation tools such as HootSuite[3] and twitterfeed[4] that allow users to pre-schedule tweets at specific intervals. These tools are not exclusive to spammers, as indicated by the presence in the random sample, though typical users are far more likely to interface with Twitter directly through the web. Interestingly, application usage amongst compromised accounts and a random sample are similar, supporting our claim that the majority of accounts that pass both automation tests are regular Twitter accounts that have been compromised.

Given our belief the majority of accounts are non-career spammers, we analyze anomalous application usage to identify instances of unauthorized third party access. For typical users, we expect tweets to originate from an array of desktop and phone applications, while spam tweets should appear from an independent application controlled by spammers. To identify this anomaly, we measure the frequency that an application is used to generate spam versus non-spam tweets on a per account basis. On average, 22% of accounts contain spam tweets that originate from applications that are *never* used for non-spam tweets. This pattern of unauthorized third party access further demonstrates that stolen Twitter accounts are being compromised and abused by spammers.

[3]http://hootsuite.com/
[4]http://twitterfeed.com/

5. SPAM CAMPAIGNS

To aid in the propagation of products and malware, spammers manage multiple accounts in order to garner a wider audience, withstand account suspension, and in general increase the volume of messages sent. To understand the collusion of accounts towards advertising a single spam website, we develop a technique that clusters accounts into *campaigns* based on blacklisted landing pages advertised by each account. We define a campaign as the set of accounts that spam at least one blacklisted landing page in common. While at least 80% of campaigns we identify consist of a single account and landing page, we present an analysis of the remaining campaigns including the number of websites hosting spam content for the campaign and number of actors involved.

5.1 Clustering URLs into campaigns

To cluster accounts into campaigns, we first define a campaign as a binary feature vector $\mathbf{c} = \{0,1\}^n$, where 1 indicates a landing page is present in the campaign and n is the total number of landing pages in our data set. When generating the feature vector for a campaign, we intentionally consider the full URL of a landing page and not its host name to allow for distinct campaigns that operate within the same domain space, such as on free web hosting, to remain separate.

Clustering begins by aggregating all of the blacklisted landing pages posted by an account i and converting them into a campaign $\mathbf{c_i}$, where each account is initially considered part of a unique campaign. Campaigns are clustered if for distinct accounts i, j the intersect $\mathbf{c_i} \cap \mathbf{c_j} \neq \emptyset$, indicating at least one link is shared by both accounts. The resulting clustered campaign $\mathbf{c_{(i,j)}} = \mathbf{c_i} \cup \mathbf{c_j}$. This process repeats until the intersection of all pairs of campaigns $\mathbf{c_i}, \mathbf{c_j}$ is empty. Once complete, clustering returns the set of landing pages for each campaign as well as the accounts participating in each campaign.

Due to our use of Twitter exclusively to identify campaigns, there are a number of limitations worth noting. First, if an account participates in multiple campaigns, the algorithm will automatically group the campaigns into a single superset. This occurs when an account is shared by two spammers, used for multiple campaigns over time by a single spammer, or compromised by different services. Alternatively, if each landing page advertised by a spammer is unique to each account, our algorithm has no means of identifying collusion and results in partitioning the campaign into multiple disjoint subsets.

5.2 Clustering results

The results of running our clustering technique on the accounts flagged by each blacklist are shown in Table 3. If there were an absence of accounts that tweet multiple scam pages, our clustering technique would return the maximum possible number of campaigns, where each landing page is considered a separate campaign. In practice this is not the case; we are able to identify multiple in-

Cluster Statistic	Google	Joewein	URIBL
Maximum possible campaigns	6,210	3,435	383,317
Campaigns identified	2,124	1,204	59,987
Campaigns with more than one account	14.50%	20%	11.46%
Campaigns with more than one page	13.09%	18.36%	27.18%

Table 3: Campaign statistics after clustering

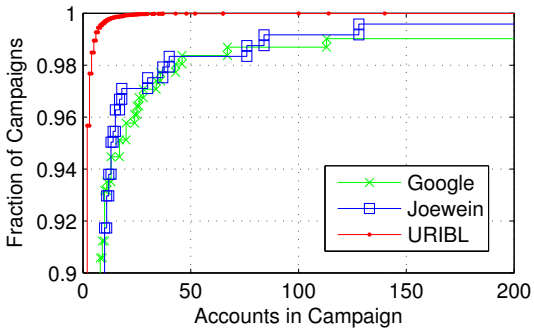


Figure 4: Number of accounts colluding in campaigns

stances where spam advertised by a group of accounts span a number of distinct landing pages, and even domains.

Analyzing the membership of campaigns, we find that at least 10% of campaigns consist of more than one account. The membership breakdown of these campaigns is shown in Figure 4. Diversity of landing pages within campaigns is slightly more frequent, as shown in Figure 5, where the use of affiliate links and multiple domains results in a greater volume of links that comprise a single campaign. While the vast majority of accounts do not collude with other Twitter members, there are a number of interesting campaigns at the tail end of these distributions that clustering helps to identify.

5.2.1 *Phishing for followers*

A particularly interesting phishing campaign that appeared during our monitoring period is websites purporting to provide victims with followers if they revealed their account credentials. In practice, these accounts are then used in a pyramid scheme to attract new victims and advertise other services.

Clustering returned a set of a 21,284 accounts that tweeted any one of 1,210 URLs associated with the campaign. These URLs directed to 12 different domains, while the full URL paths contained affiliate information to keep track of active participants. To understand spamming behavior within the campaign, we fractured users into *subcampaigns*, where a subcampaign is a set of users that share identical feature vectors, rather than the original criteria of sharing at least one link in common. From the initial campaign, hundreds of subcampaigns appear. Of the 12 distinct domains, each has a independent subcampaign consisting of on average 1,400 participants, accounting for roughly 80% of the original campaign members. The remaining 20% of participants fall into multiple clusters due to signing up for multiple follower services, accounting for why the independent campaigns were initially merged.

Defining features. This campaign makes up a significant portion of the tweets flagged by the Google blacklist, and shows surprisingly large user involvement and frequent tweeting. Using the χ^2 and entropy tests, we find that a large fraction of the users, 88% in our set, tweeting for this campaign are compromised users, adding to

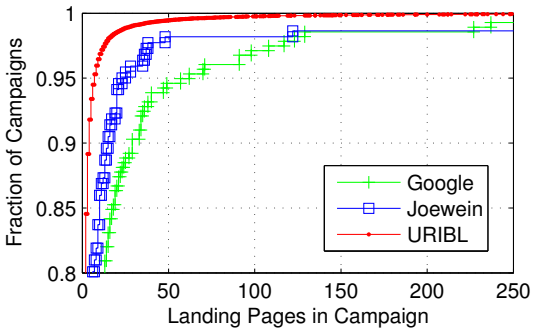


Figure 5: Number of landing pages targeted by campaigns

the evidence that phished accounts are used to further promote the phishing campaign. A defining feature of tweets in this campaign is the extensive use of hashtags, 73% of the tweets sent contained a hashtag. Hash tags are frequently reused and typically denote the subcampaign (such as #maisfollowers). For the URLs being tweeted, most have a redirect chain consisting of a single hop, from a shortened URL to the landing page, though affiliate tracking typically introduces a second hop (shortened URL -> affiliate link -> landing page). In some cases, the landing page itself appears in tweets. We have also observed that the phishing sites plainly advertise the service to get more followers.

5.2.2 *Personalized mentions*

Of the campaigns we identify as spam, one particular campaign run by *http://twitprize.com* uses Twitter to draw in traffic using thousands of career accounts that exclusively generate spam telling users they had won a prize. Clustering returns a set of 1,850 accounts and 2,552 distinct affiliate URLs that were all shortened with tinyurl. Spam within the campaign would target victims by using mentions and crafting URLs to include the victim's Twitter account name to allow for personalized greetings. Promising a prize, the spam page would take a victims address information, require a survey, list multiple mailers to sign up for, and finally request the user either sign up for a credit card or subscribe to a service.

Defining features. This campaign is dominated by tweet URLs from tinyurl pointing to unique, victim-specific, landing pages at *http://twitprize.com* with no intermediate redirects. Of the tweets containing URLs in this campaign, 99% are a retweet or mention. The heavy use of usernames in tweets is an interesting characteristic, unique to this type of campaign. Unlike the previous phishing campaign, we find infrequent use of hashtags, with only 2% of tweets containing a hashtag. The accounts that tweet URLs in this campaign pass the entropy tests since each tweet contains a different username and the links point to distinct twitprize URLs. Of the accounts participating, 25% have since been suspended by Twitter.

5.2.3 *Buying retweets*

One of the primary challenges for spammers on Twitter is to gain a massive following in order to increase the volume of users that will see a spam tweet. To circumvent this challenge, a number of services have appeared that sell access to followers. One such service, *retweet.it*, purports to retweet a message 50 times to 2,500 Twitter followers for $5 or 300 times to 15,000 followers for $30. The accounts used to retweet are other Twitter members (or bots) who sign up for the retweet service, allowing their accounts to be used to generate traffic.

Defining features. While the service itself does not appear to be a

scam, it has been employed by spammers. Using a unique feature present in all *retweet.it* posts to generate a cluster, we identify 55 accounts that retweeted a spam post soliciting both malware and scams. The χ^2 test indicate that 84% of the accounts are career spammers.

5.2.4 Distributing malware

Using clustering, we identified the largest campaign pushing malware in our data set, consisting 113 accounts used to propagate 57 distinct malware URLs. The content of the sites include programs that bring satellite channels to a computer that are "100% adware and spyware free" and an assortment of other scams. In addition to serving malware, some sites advertised by the campaign were reported by Google's blacklist for drive by downloads.

Defining features. The top malware campaign is significantly different than other campaigns, with a relatively small account base and few tweets. The accounts that tweet links in this cluster tend to be career spammers, indicating that the malware is not compromising Twitter accounts in order to self propagate, a feature found among Twitter phishing URLs. One difference from other campaigns is this use of redirects to mask the landing page. Since both Twitter and shortening services such as *bit.ly* use the Google Safebrowsing API to filter links, if a *bit.ly* URL is to be placed in tweets, the redirect chain must at least be two hops (*bit.ly* → intermediate → malware landing site). Two hops is not enough though, as the Safebrowsing list contains both sites that serve as well as sites that redirect to malware, requiring at least an additional hop to be used to mask it from initial filtering.

5.2.5 Nested URL shortening

In addition to locating large campaigns, clustering helps to identify instances of URL compression where multiple links posted in tweets all resolve to the same page. One such campaign consisted of 14 accounts soliciting a financial scam. While unremarkable for its size, the campaign stands out for its use of multiple redirector services, totaling 8 distinct shortening domains that appear in tweets. In turn, each initial link triggers a long chain of nested redirects that leads our crawler through *is.gd* → *short.to* → *bit.ly* before finally resolving to the scam page. While the motivation for nested redirects is unclear, it may be a result of spam filtering done on the part of shortening services. By nesting URLs, filtering based on domains or full URLs is rendered obsolete less the final URL is resolved, which we discuss further in Section 6

6. BLACKLIST PERFORMANCE

Given the prevalence of spam throughout Twitter, we examine the degree to which blacklists could stem the spread of unsolicited messages. Currently, Twitter relies on Google's SafeBrowsing API to block malicious links, but this filtering only suppresses links that are blacklisted at the time of its posting; Twitter does not retroactively blacklist links, allowing previously undetected malicious URLs to persist. To measure how many tweets slip through Twitter's defenses, and whether the same would be true for URIBL and Joewein, we examine a number of blacklist characteristics, including delay, susceptibility to evasion, and limitations that result if we restrict filtering to considering only domains rather than the full paths of spam websites.

6.1 Blacklist delay

Using historical data for the URIBL, Joewein, and Google blacklists, we can measure the delay between a tweet's posting and the time of its subsequent blacklisting. For cases where a spam URL embedded in a tweet appeared on a blacklist prior to appearing on Twitter, we say that the blacklist *leads* Twitter. Conversely, a blacklist *lags* Twitter if posted URLs reach the public before becoming blacklisted. Lead and lag times play an important role in determining the efficiency of blacklists. For example, for long lag periods spam filters must maintain a large index of URLs in stale tweets to retroactively locate spam. Furthermore, depending on the rate at which users click on spam links, long lag periods can result in little protection unless spammers reuse links even after they appear on blacklists.

Link Statistics	URIBL	Joewein	Google Malware	Google Phishing
Flagged before posting	27.17%	3.39%	7.56%	1.71%
Flagged after posting	72.83%	96.61%	92.44%	98.29%
Avg. lead period (days)	29.40	13.41	29.58	2.57
Avg. lag period (days)	-21.93	-4.28	-24.90	-9.01
Overall avg. (days)	-12.70	-3.67	-20.77	-8.82

Table 4: Blacklist performance, measured by the number of tweets posted that lead or lag detection. Positive numbers indicate lead, negative numbers indicate lag.

Link Statistics	URIBL	Joewein	Google Malware	Google Phishing
Flagged before posting	50.19%	20.31%	18.89%	15.38%
Flagged after posting	49.81%	79.69%	81.11%	84.62%
Avg. lead period (days)	50.53	15.51	28.85	2.50
Avg. lag period (days)	-32.10	-5.41	-21.63	-10.48
Overall avg. (days)	9.36	-1.16	-12.10	-8.49
Total domains flagged	1620	128	625	13

Table 5: Blacklist performance, measured by lead and lag times for unique domains posted.

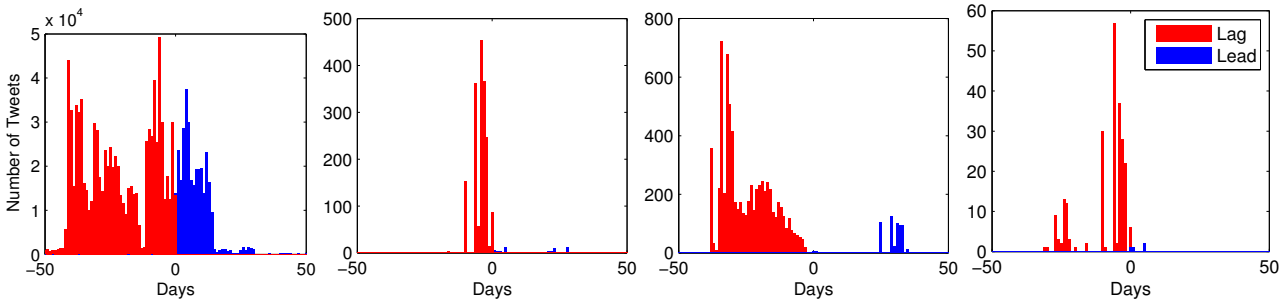
We begin measuring blacklist delay by gathering the timestamps for each tweet of a blacklisted URL. For URLs spammed in multiple tweets, we consider each posting as a unique, independent event. Table 4 shows the lead and lag times for tweets, where we see that the majority of spam tweets appear on Twitter multiple days prior to being flagged in blacklists, and in the case of URIBL and Google, multiple weeks. A more extensive presentation of blacklist delay can be seen in Figure 6, showing the volume of tweets per lead and lag day. It is important to note that Twitter use of Google's Safebrowsing API to filter links prior to their posting biases our analysis towards those links that pass through the filter, effectively masking the lead time apart from URLs that spammers obfuscated with shorteners to avoid blacklisting.

Table 5 shows the same lead and lag periods but weighted by unique domains rather than by individual tweets. While blacklisting timeliness improves from this perspective, this also indicates that domains previously identified as spam are less likely to be reposted, limiting the effectiveness of blacklisting.

To understand the exposure of users due to blacklist lag, we measured the rate that clicks arrived for spam links. Using daily clickthrough data for a random sample of 20,000 spam links shortened with *bitly*, we found that 80% of clicks occur within the first day of a spam URL appearing on Twitter, and 90% of clicks within the first two days. Thus, for blacklisting to be effective in the context of social networks, lag time must be effectively zero in order to prevent numerous users from clicking on harmful links.

6.2 Evading blacklists

The success of blacklists hinges on the reuse of spam domains; if every email or tweet contained a unique domain, blacklists would



(a) URIBL (b) Joewein (c) Google - Malware (d) Google - Phishing

Figure 6: Volume of spam tweets encountered, categorized by either lagging or leading blacklist detection

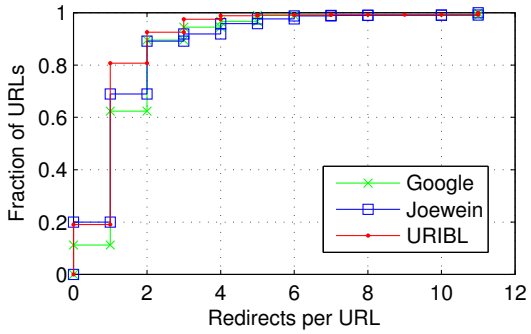


Figure 7: Frequency of redirects and nested redirects amongst distinct spam URLs

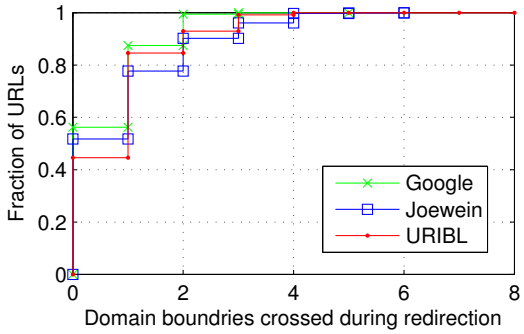


Figure 8: Frequency of cross-domain redirects amongst distinct spam URLs containing at least one hop

be completely ineffective. While the registration of new domains carries a potentially prohibitive cost, URL shortening services such as *bitly*, *tinyurl*, *is.gd*, and *ow.ly* provide spam orchestrators with a convenient and free tool to obfuscate their domains.

By following shortened URLs, we found over 80% of distinct links contained at least one redirect, as shown in a breakdown in Figure 7. In particular, redirects pose a threat to blacklisting services when they cross a domain boundary, causing a link to appear from a non-blacklisted site as opposed to a blacklisted landing page. Figure 8 shows the cross-domain breakdown for distinct URLs seen in links containing at least one redirect. Roughly 55% of blacklisted URLs cross a domain boundary.

The effect of shortening on Twitter's malware defenses (filtering via Google's Safebrowsing API) appears quite clearly in our data set. Disregarding blacklist delay time, 39% of distinct malware and phishing URLs evade detection via use of shorteners. Despite the small fraction, these links make up *over 98% of malicious tweets* identified by our system. Even in the event a shortened URL becomes blacklisted, generating a new URL comes at effectively no cost. Without the use of crawling to resolve shortened URLs, blacklists become much less effective.

6.3 Domain blacklist limitations

For blacklists based only on domains rather than full URLs, such as URIBL and Joewein, false positives pose a threat of blacklisting entire sites. Looking through the history of URIBL and Joewein, we identified multiple mainstream domains that were blacklisted prior to our study, including *ow.ly*, *tumblr*, and *friendfeed*. Each of these services allow users to upload content, giving rise to the potential for abuse by spammers.

The presence of user-generated content and mashup pages presents a unique challenge for domain blacklists. For instance, while *ow.ly* merely acts as a redirector, the site embeds any spam pages to which it redirects in an iFrame, causing a browser's address bar to always display *ow.ly*, not the spam domain. When faced with mashup content, individual cross-domain components that make up a page must be blacklisted rather than the domain hosting the composite mashup. This same challenge exists for Web 2.0 media where content contributed by users can affect whether a domain becomes blacklisted as spam. For *tumblr* and *friendfeed*, we identified multiple cases in our data set where the domains were used by spammers, but the majority of accounts belong to legitimate users. The appearance and subsequent deletion of social media domains within URIBL and Joewein disguises the fact that the domains are being abused by spammers. To address the issue of spam in social media, individual services can either be left to tackle the sources of spam within their own sites, or new blacklists must be developed akin to Google's Safebrowsing API that go beyond domains and allow for fine-grained blacklisting.

7. CONCLUSION

This paper presents the first study of spam on Twitter including spam behavior, clickthrough, and the effectiveness of blacklists to prevent spam propagation. Using over 400 million messages and 25 million URLs from public Twitter data, we find that 8% of distinct Twitter links point to spam. Of these links, 5% direct to malware

and phishing, while the remaining 95% target scams. Analyzing the account behavior of spammers, we find that only 16% of spam accounts are clearly automated bots, while the remaining 84% appear to be compromised accounts being puppeteered by spammers. Even with a partial view of tweets sent each day, we identify coordination between thousands of accounts posting different obfuscated URLs that all redirect to the same spam landing page. By measuring the clickthrough of these campaigns, we find that Twitter spam is far more successful at coercing users into clicking on spam URLs than email, with an overall clickthrough rate of 0.13%.

Finally, by measuring the delay before blacklists mark Twitter URLs as spam, we have shown that if blacklists were integrated into Twitter, they would protect only a minority of users. Furthermore, the extensive use of URL shortening services masks known-bad URLs, effectively negating any potential benefit of blacklists. We directly witness this effect on Twitter's malware and phishing protection, where even if URLs direct to sites known to be hostile, URL shortening allows the link to evade Twitter's filtering. To improve defenses for Twitter spam, URLs posted to the site must be crawled to unravel potentially long chains of redirects, using the final landing page for blacklisting. While blacklist delay remains an unsolved challenge, retroactive blacklisting would allow Twitter to suspend accounts that are used to spam for long periods, forcing spammers to obtain new accounts and new followers, a potentially prohibitive cost.

8. REFERENCES

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

[2] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the 4th International Conference on Weblogs and Social Media*, 2010.

[3] A. Chowdhury. State of Twitter spam. http://blog.twitter.com/2010/03/state-of-twitter-spam.html, March 2010.

[4] F-Secure. Twitter now filtering malicious URLs. http://www.f-secure.com/weblog/archives/00001745.html, 2009.

[5] R. Flores. The real face of Koobface. http://blog.trendmicro.com/the-real-face-of-koobface/, August 2009.

[6] Google. Google safebrowsing API. http://code.google.com/apis/safebrowsing/, 2010.

[7] D. Harvey. Trust and safety. http://blog.twitter.com/2010/03/trust-and-safety.html, March 2010.

[8] D. Ionescu. Twitter Warns of New Phishing Scam. http://www.pcworld.com/article/174660/twitter_warns_of_new_phishing_scam.html, October 2009.

[9] D. Irani, S. Webb, and C. Pu. Study of static classification of social spam profiles in MySpace. In *Proceedings of the 4th International Conference on Weblogs and Social Media*, 2010.

[10] J. John, A. Moshchuk, S. Gribble, and A. Krishnamurthy. Studying spamming botnets using Botlab. In *Usenix Symposium on Networked Systems Design and Implementation (NSDI)*, 2009.

[11] 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*, pages 3–14. ACM, 2008.

[12] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the International World Wide Web Conference*, 2010.

[13] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: Social honeypots + machine learning. In *Proceeding of the SIGIR conference on Research and Development in Information Retrieval*, pages 435–442, 2010.

[14] R. McMillan. Stolen Twitter accounts can fetch $1,000. http://www.computerworld.com/s/article/9150001/Stolen_Twitter_accounts_can_fetch_1_000, 2010.

[15] B. Meeder, J. Tam, P. G. Kelley, and L. F. Cranor. RT @IWantPrivacy: Widespread violation of privacy settings in the Twitter social network. In *Web 2.0 Security and Privacy*, 2010.

[16] J. O'Dell. Twitter hits 2 billion tweets per month. http://mashable.com/2010/06/08/twitter-hits-2-billion-tweets-per-month/, June 2010.

[17] A. Pitsillidis, K. Levchenko, C. Kreibich, C. Kanich, G. Voelker, V. Paxson, N. Weaver, and S. Savage. Botnet Judo: Fighting spam with itself. 2010.

[18] Z. Qian, Z. Mao, Y. Xie, and F. Yu. On network-level clusters for spam detection. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2010.

[19] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 workshop*. Madison, Wisconsin: AAAI Technical Report WS-98-05, 1998.

[20] E. Schonfeld. When it comes to URL shoteners, bit.ly is now the biggest. http://techcrunch.com/2009/05/07/when-it-comes-to-url-shorteners-bitly-is-now-the-biggest/, May 2009.

[21] K. Thomas and D. M. Nicol. The Koobface botnet and the rise of social malware. Technical report, University of Illinois at Urbana-Champaign, July 2010. https://www.ideals.illinois.edu/handle/2142/16598.

[22] Twitter. The Twitter rules. http://help.twitter.com/forums/26257/entries/18311, 2009.

[23] URIBL. URIBL.COM – realtime URI blacklist. http://uribl.com/, 2010.

[24] Y. Wang, M. Ma, Y. Niu, and H. Chen. Spam double-funnel: Connecting web spammers with advertisers. In *Proceedings of the International World Wide Web Conference*, pages 291–300, 2007.

[25] J. Wein. Joewein.de LLC – fighting spam and scams on the Internet. http://www.joewein.net/.

[26] C. Wisniewski. Twitter hack demonstrates the power of weak passwords. http://www.sophos.com/blogs/chetw/g/2010/03/07/twitter-hack-demonstrates-power-weak-passwords/, March 2010.

[27] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming botnets: Signatures and characteristics. *Proceedings of ACM SIGCOMM*, 2008.