Click Traffic Analysis of Short URL Spam on Twitter

De Wang, Shamkant B. Navathe, Ling Liu, Danesh Irani, Acar Tamersoy, Calton Pu
College of Computing
Georgia Institute of Technology
Atlanta, Georgia 30332, United States
Email: {wang6, sham, lingliu, danesh, acar.tamersoy, calton.pu}@cc.gatech.edu

***Abstract*—With an average of 80% length reduction, the URL shorteners have become the norm for sharing URLs on Twitter, mainly due to the 140-character limit per message. Unfortunately, spammers have also adopted the URL shorteners to camouflage and improve the user click-through of their spam URLs. In this paper, we measure the misuse of the short URLs and analyze the characteristics of the spam and non-spam short URLs. We utilize these measurements to enable the detection of spam short URLs. To achieve this, we collected short URLs from Twitter and retrieved their click traffic data from Bitly, a popular URL shortening system. We first investigate the creators of over 600,000 Bitly short URLs to characterize short URL spammers. We then analyze the click traffic generated from various countries and referrers, and determine the top click sources for spam and non-spam short URLs. Our results show that the majority of the clicks are from direct sources and that the spammers utilize popular websites to attract more attention by cross-posting the links. We then use the click traffic data to classify the short URLs into spam vs. non-spam and compare the performance of the selected classifiers on the dataset. We determine that the Random Tree algorithm achieves the best performance with an accuracy of 90.81% and an F-measure value of 0.913.**

***Keywords*—traffic analysis, short URL spam, Twitter**

I. Introduction

Social networks attract millions of users who want to share information and connect with people. Twitter, a popular social network, has over 400 million members and it allows them to post 140-character tweets (messages) to their network of followers. Given the limited length of a tweet, URL shorteners have quickly become the de facto method to share links on Twitter [1].

Twitter, due to its large audience and information reach, attracts spammers [2], [3], [4], [5], [1], [6], [7], [8]. Even though spammers have limited flexibility with the 140-character limit for a tweet, they utilize URL shorteners to camouflage their spam links [9], [10], [11], [12]. This enables spammers to hide the true domain of the URL, thereby might prevent Twitter from effectively applying blacklists to filter out such spam.

The popular URL shortener websites such as Bit.ly (henceforth referred to as Bitly) provide interfaces that allow users to convert long URLs into short URLs [13], [14]. After receiving a long URL, the services typically use a hash function to map the long URL to a short string of alphanumeric characters, which is then appended to the domain name of the shortener and returned as the short URL. For instance, the long URL http://www.google.com might be shortened as http://bit.ly/olDmsz. The hash function takes into account several factors, such as whether the long URL has already been mapped to a short URL. In this case, the shorteners typically return the existing short URL rather than generating a new one for the input long URL.

In this paper, we perform an analysis on short URL spam by investigating their click traffic with the following goals. First, we aim to determine the feasibility of efficiently collecting the click traffic of short URLs. This is important because a social network typically contains a massive number of short URLs and an efficient mechanism is needed to collect their click traffic. Second, we aim to discover significant patterns in the click traffic of a given set of spam short URLs. Third, we aim to determine the feasibility of detecting short URL spam effectively. This is particularly important because spam can lead to loss and damage [15], [16].

The highlights of our work can be summarized as follows:

- We generate a large-scale click traffic dataset for short URL spam;
- We obtain several findings about short URL spam through an in-depth analysis of creators and click sources of short URLs;
- We demonstrate the feasibility of detecting short URL spam by classification based on the click traffic features.

The remainder of the paper is organized as follows. We motivate the problem further in Section II. Section III introduces the approach developed for collecting the short URLs and the datasets used in the experiments. Section IV provides the results of the statistical analysis of the short URLs. Section V describes our approach of classifying short URLs based on their click traffic features. Section VI presents the evaluation metrics and classification results using different classifiers. We survey the related work in Section VII and conclude the paper in Section VIII.

II. Motivation

Existing studies have focused on URL spam detection and revolved primarily around blacklists and domain reputation. Blacklists are typically built for previously-classified URLs, domains, or IP addresses, and incoming URLs are simply checked against them [17]. These techniques do not work

effectively when spammers employ short URLs. This is because blacklists based on domains and IP addresses incorrectly flag the short URL generated by the URL shortening service instead of the long, malicious URL behind by the short URL, and furthermore, spammers generate a new short URL as soon as the previous one is blacklisted. One solution to this problem might be to resolve each shortened URL and fetch the web page associated with it. Previous studies [18], [19] on web page content classification have shown high accuracy, however these techniques, although highly accurate, result in high classification cost and incur significant delay due to the fact that they need to download the content. Additionally, these techniques do not work for some malware customized web pages that are capable of dynamically changing content to confuse the content-based spam filters.

Similar to traffic analysis approach used in network anomaly detection [20], we aim to investigate short URL click traffic in order to detect patterns for short URL spam. Also, Las-Casas et al. [21] have efficiently used network metrics to detect spammers at the source network instead of content. In this paper, we assume that spammers propagate spam URL in different way from legitimate users and that it should be less probable for people choosing to click spam URL. Next, we describe how we obtained the short URLs and their click traffic using public APIs.

III. Data Collection

In this section, we first describe the approach used for data collection and the properties of the datasets. And then, we discuss the ground truth labeling of the datasets.

A. Collection Approach

To collect data, we use two APIs: Twitter APIs [22] and Bitly APIs [23]. Twitter APIs provide two types of objects: a user profile and a tweet. We extract URL links from the tweets and filter out all the short URLs. Bitly APIs provide four major types of meta-data for each short URL: *info*, *clicks*, *countries*, and *referrers*. *Info* contains the properties of the short URL, such as the long URL behind the short URL. *Clicks* contains the total amount of clicks for the short URL. *Countries* and *referrers* record the number of clicks from various countries and referrers, respectively. Here, "referrers" correspond to the applications or web pages that contain the short URLs.

B. Datasets

The details of our datasets are as follows:

Twitter Dataset: We collected data of over 900,000 Twitter users, about over 2.4 million tweets, fetching any links in the tweets. The tweets were gathered by querying the top trending topics every minute and they represent about 600 topics over the span of November 2009 to February 2010. Twitter users or tweets marked as suspended or removed due to terms of service violations are explicitly marked as spam in the dataset. There are over 26,000 such users and 138,000 such tweets.

We extracted all the short URLs from the Twitter dataset and then ranked the short URL providers based on the total number of URLs created by the providers. The result is shown in Table I.

TABLE I. Top-10 short URL providers in the Twitter dataset

Short URL Provider	*Count*
Bit.ly	641,423
t.co	129,677
Tiny.com	62,488
Ow.ly	42,838
Is.gd	14,664
Goo.gl	13,122
j.mp	8,963
Su.pr	3,764
Twurl.nl	2,807
Migre.me	2,788

We observed that Bitly generated the majority of the short URLs in the dataset, achieving about 57% of the total URLs. As Bitly also has public APIs that enabled us to download click traffic meta-data of the short URLs, we decide to focus on the Bitly short URLs and generate the Bitly dataset as follows.

Bitly Dataset: We extracted all the Bitly short URLs from the Twitter dataset and fetched their click traffic using the Bitly APIs. The click traffic dataset consists of four types of meta-data:

- **Basic information** containing five attributes: id (identification number of the short URL), url (short URL address), long_url (URL that the short URL points to), title (title of the web page), created_by (creator of the short URL);
- **Number of user-level and global clicks** containing three attributes: url_id (identification number of the short URL), user_clicks (total number of clicks received), global_clicks (total number of clicks received globally[1]);
- **Country click distribution** containing three attributes: url_id (identification number of the short URL), countries (list of countries), and clicks (number of clicks from the countries);
- and **Referrer click distribution** containing three attributes: url_id (identification number of the short URL), referrers (web pages or applications that referred the short URL), clicks (number of clicks from the referrers). The total number of short URLs is $641,423$, including $18,496$ spam short URLs and $622,927$ legitimate short URLs.

We have to admit that Bitly is no longer the default link shortener on Twitter [24]. However, our work is independent to the shortening service provider since almost all shortening service providers could provide the similar information above for each short URL. Therefore, our research methods could be adapted easily by any other shortening URL providers. Next, we explain the labeling of the short URLs.

C. Data Labeling

As mentioned earlier, the tweets marked as removed due to terms of service violations are explicitly marked as spam in the Twitter dataset. We utilized this information as ground

[1]There may be multiple short URLs pointing to the same long URL. This attribute records the total number of clicks received for all the short URLs pointing to the same long URL.

truth and assumed that these tweets contain malicious content, hence we labeled them as spam tweets.

The average account suspension rate is about 3% in the Twitter dataset. Account suspension by creation date is shown in Figure 1.

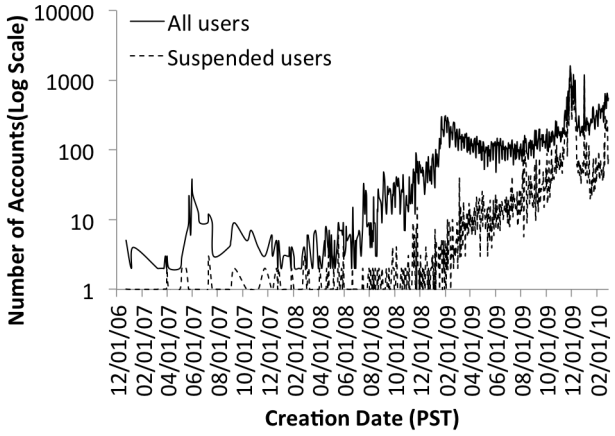


Fig. 1. Account suspension by creation time in the Twitter dataset (time unit = day)

Based on previous work which has shown that URL links in spam messages have a high probability to be spam URLs [25], we labeled the short URLs in the spam tweets as spam short URLs.

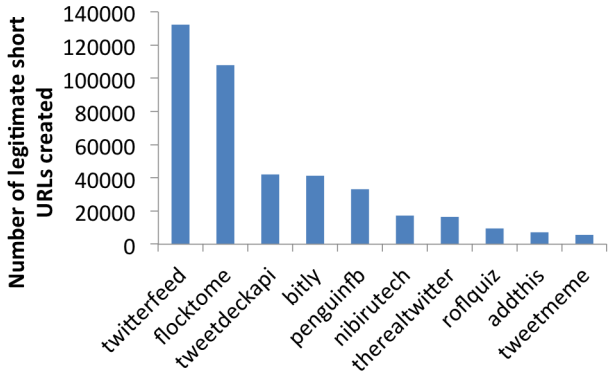
We also checked the short URLs (including the final URLs and URLs in the redirection chain) against several public blacklists to validate the ground truth. The public blacklists included Google Safe Browsing, McAfee SiteAdvisor, URIBL, SURBL, and Spamhaus [17], [26], [27], [28], [29]. Google Safe Browsing allows users to query URLs against Google's constantly updated lists of phishing and malware pages. McAfee SiteAdvisor provides safety test results for the websites and shows a warning when the URL links to spam. URIBL, SURBL and Spamhaus are using similar mechanisms; they all contain suspicious websites appeared in spam emails. If the URLs were listed in any of the blacklists, we labeled them as spam. Since there was a delay between the time we generated the dataset and the time we labeled the dataset, the lag effect of blacklist validation was not a problem.

IV. Data Analysis

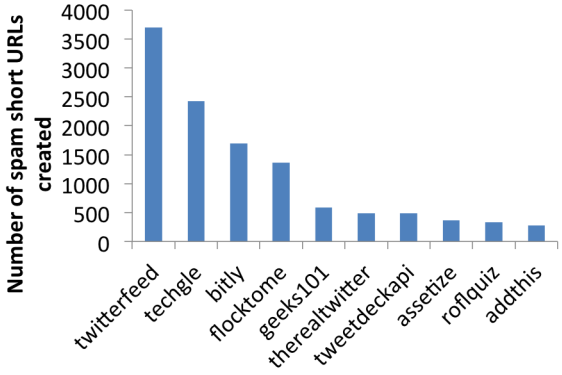
In this section, we start with our analysis by focusing on the various attributes in the Bitly dataset, performing a statistical creator analysis and click source analysis.

A. Creator Analysis

We first focus on the creators of the short URLs. On Bitly, a creator is either a regular user or an enterprise user who generates the short URL for its services [30]. Through the creator analysis, we try to find out whether it could reveal the real spammers behind the scene. In our dataset, the total number of creators are $32,452$. Figure 2(a) depicts the top-10 creators of legitimate short URLs. Of these creators, *twitterfeed* is a utility allowed to feed content to Twitter, *tweetdeckapi* is the API service for tweetdeck, which is a social media dashboard application for management of Twitter, *penguinfb* is a service for sending status updates to Twitter from Facebook, *niburutech* is the software company behind a widely used Twitter client, *roflquiz* is a website offering funny Twitter quizzes, *addthis* is a social bookmarking service for Twitter, and *tweetmeme* is a service that determines the popular links on Twitter. We were not able to obtain accurate information about *flocktome* and *therealtwitter*. Similarly, Figure 2(b) depicts the top-10 creators of the spam short URLs. Of these creators, *assetize* is an advertising network for Twitter. We were not able to obtain accurate information about *techgle* and *geeks101*. We further observe that the total number of spam URLs created by the top-3 legitimate creators (i.e., twitterfeed, bitly, and tweetdeckapi) accounts for more than 31% of all the spam short URLs in the dataset.



(a) Top-10 creators of legitimate short URLs



(b) Top-10 creators of spam short URLs

Fig. 2. Creators of the short URLs

We subsequently computed the percentage of the spam short URLs created by each creator. We observe that the number of creators who have created 80% or more of short URL spam in all URLs is 344. This corresponds to over 1% of the creator population. The total number of short URLs created by these creators corresponds to more than 31% of all the spam short URLs in the dataset.

Also, we notice that some creators that are assumed to be legitimate (e.g., twitterfeed) created spam short URLs. One reason for this might be the fact that these legitimate creators are not individual creators in the sense that they automatically shorten the URLs posted by the Twitter users. Moreover, it tells

us that we cannot determine whether the creator is a spammer based on spam URLs they may create when the creator is an enterprise user. However, If the creator is individual user who generates spam URLs, we could track it back through the URLs and block it away. Furthermore, if the enterprise user has the mapping record which indicates who is the original creator for the spam URL, the enterprise user could cooperate with shortening service provider to lock down the culprit.

Another thought about creators is that whether we could determine that they are spammers by calculating out the percentage of spam URLs in all URLs that have been created by particular creators. Generally speaking, we believe that legitimate users should create more legitimate URLs than spam URLs. Therefore, we rank the creators based on the percentage of spam URLs and the total number of spam URLs they created in the decreasing order.

TABLE II. TOP-10 CREATORS THAT CREATED ONLY SPAM SHORT URLS

Creator	*Spam URLs / All URLs*
dailypiff187	150/150
golfhonchonow	72/72
headlinehoncho	63/63
newswatchphilly	56/56
mskaya4u	56/56
golfhonchotoo	50/50
golfhoncho	48/48
breakingnewssource	47/47
onlinenewsblast	47/47
portlandtimestribune	46/46

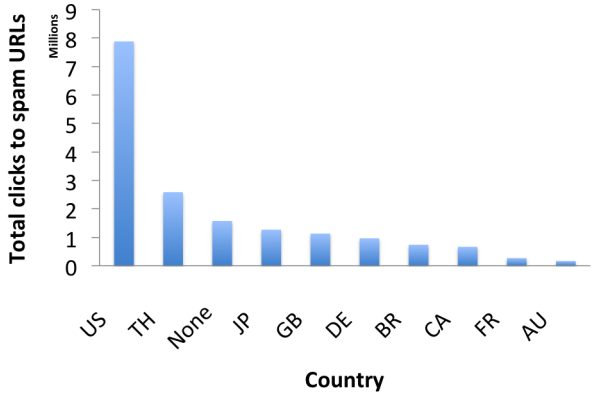
Table II lists the top-10 creators that created only spam short URLs, sorted by the total number of short URLs created. We observe that the short URLs created by these 10 creators account for more than 3.4% of all the spam short URLs, which is a significant portion considering the total number of creators in the dataset. We were not able to obtain additional information about these creators from Bitly in order to decide whether all the short URLs they created are spam or not, nonetheless we believe that this kind of ranking is useful to classify whether a user of a URL shortening service is a spammer or not.

Meanwhile, prior research [31] has concluded that Twitter is mostly a news propagation network, with more than 85% of trending topics reflecting headline news. It is also interesting to see that many creators in Table II have news-related names (e.g., *newswatchphilly*, *breakingnewssource* and *onlinenewsblast*), showing that spammers are likely using this fact in order to increase their success rates. In addition, it means that spammers may employ URL shortening services by registering as enterprise users with trustworthy business names, which is really hard for those URL shortening service providers to distinguish them from legitimate enterprise users.

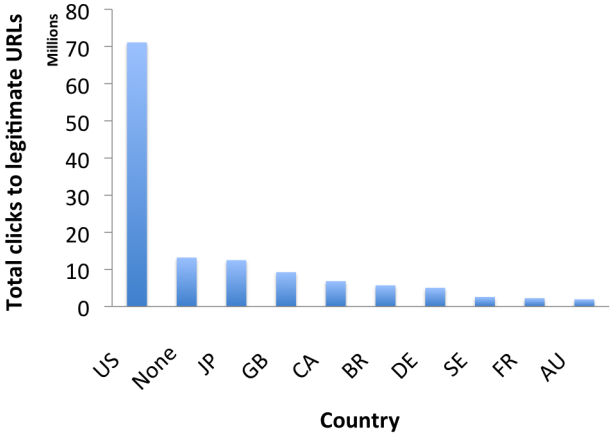
B. Click Source Analysis

In addition to creator analysis, we also investigated the sources of click traffic which may shed light on special patterns in click traffic of spam URLs. There are two types of click sources available in the Bitly dataset, namely the country click source and the referrer click source. Each short URL is associated with a country list and a referrer list, which contain click distributions coming from the countries and referrers, respectively. For example, short URL http://bit.ly/olDmsz has country list (US: 9 clicks) and referrer list (direct (email, IM, apps): 8 clicks and bitly.com: 1 click). This short URL is created for the long URL http://www.google.com for test purpose.

1) Country Source Analysis: The country list of short URL tells us the clicks from each country, which will help us find out those countries which generate high click traffic for spam URLs and legitimate URLs as well. We first aggregate the country click source by country name and list the top-10 countries based on the clicks to spam URLs and legitimate URLs. The results are shown in Figure 3. Here, "None" country source means that the country source is unknown to URL shortening service provider – Bitly in this case.



(a) Top-10 countries based on clicks to spam URLs



(b) Top-10 countries based on clicks to legitimate URLs

Fig. 3. Country click sources of the short URLs

From Figures 3(a) and 3(b), we have the following observations: (i) United States (US), Japan (JP), Great Britain (GB), Germany (DE), Brazil (BR), Canada (CA), France (FR), and Australia (AU) are in the lists for both spam and legitimate URLs, (ii) Thailand (TH) is ranked the second in the list for spam URLs but it is not in the list for legitimate URLs, and (iii) Concerning the relative order of the countries, Canada swaps positions with Germany in the list for legitimate URLs compared to that for spam URLs. The others remain in the same order.

For the second observation, after checking the clicks from Thailand, we determined that the reason is the spam URL *http://www.uggdiscount.co.uk/*, for which the total number of clicks is $2,554,121$. The creator of the short URL in Bitly is *mysocial*. It shows that spam URL may generate heavy traffic using URL shortening service.

Therefore, the observations shows some differences in terms of click traffic from source countries between spam URLs and legitimate URLs. But from spam detection perspective, we also need to look into click distribution of countries for each URL. Sophisticated spammers may spread spam URLs across the world that results in that more country codes in the list. We will take the distribution into account in our classification section.

2) Referrer Source Analysis: Referrer is the web page or application that contain the link to the web page pointed by the short URL. A referrer must have generated click traffic so as to appear in the Bitly dataset. The referrer list shows how many clicks come up from each referrer after the short URL is posted on them. We aggregate the referrer click source by referrer and list the top-10 referrers of the spam URLs and legitimate URLs based on clicks in Tables III and IV, respectively. Here, "direct" referrer means that referrers such as email messages, instant messages, and apps.

TABLE III. TOP-10 REFERRERS OF SPAM URLS BASED ON CLICKS

Referrer	*Clicks*
direct	11,392,281
http://twitter.com/	2,619,560
http://twitter.com/home	229,628
http://td.partners.bit.ly	155,050
http://iconfactory.com/twitterrific	138,392
http://www.facebook.com/home.php	132,627
http://real-url.org	114,789
http://www.youtube.com/watch	105,056
http://www.facebook.com/	89,988
http://untiny.me	80,359

TABLE IV. TOP-10 REFERRERS OF LEGITIMATE URLS BASED ON CLICKS

Referrer	*Clicks*
direct	44,149,149
http://twitter.com/	10,947,917
http://td.partners.bit.ly	1,421,585
http://twitter.com/home	1,154,206
http://www.facebook.com/1.php	1,120,563
http://www.facebook.com/home.php	994,012
http://iconfactory.com/twitterrific	931,254
http://www.facebook.com/	774,080
http://twitter.com/ricky_martin	395,698
http://www.youtube.com/watch	385,082

From Tables III and IV, we make the following observations: (i) The majority of the clicks are from direct sources such as email clients, instant messages and applications, and (ii) The spammers utilize popular social media such as Twitter and Facebook for short URL spam to attract more attention.

The observations shows that short URLs are very popular not only on social media but also on other kinds of media such as traditional emails and mobile phones.The reason is that social networking sites connect those media together like the prediction that everything will be connection in twenty years [32] . Moreover, spammers should know that all those media are connected and propagate spam across them through short URLs.

The same to the country source analysis, we need to look into click distribution of referrers for each URL as well. We believe that spammers try to use different channels as many as possible. And we will take it into account in the classification section. In addition, the referrer list exposes the places where spammers post short URL spam. Thus, after short URL spam are detected, spam detection team on social media could use the referrer list to track all the places having the short URL spam except direct source since it does not or cannot provide specific addresses in the URI form.

V. CLASSIFICATION

In this section, we first introduce the features used in short URL classification . Then, we describe the data filtering process and the classifiers used in our classification framework.

A. Classification Features

As mentioned in Section IV-B, there are four types of meta-data available in the Bitly dataset. We keep most of the attributes in the tables as features and additionally add aggregate features for classification. All following features are Twitter-independent features so that our classifier could be easily adapted to detect short URL spam on any other social media.

For each short URL, we have chosen the following features for classification:

- **Clicks:** user_clicks (total number of clicks received), global_clicks (total number of clicks received globally), and the ratio between user_clicks and global_clicks;
- **Countries:** country count and features of click distribution (mean and standard derivation);
- **Referrers:** referrer count and features of click distribution (mean and standard derivation).

Clicks features provide us the quantitative measure of click traffic over the lifetime of short URLs in big picture. Countries features show us click distribution from source countries and referrers features give us click distribution from source referrers. We use those features to test our assumption that spammers are propagating spam across multiple countries using many referrers as they can. Also, we try to find out which feature could express the most discriminative power in short URL classification.

B. Machine Learning Classifiers

We use the various classifiers implemented in the Weka software package [33]. Weka is an open source collection of machine learning algorithms and has become the standard tool in the machine learning community. The classifiers used in the classification framework include Random Forest, Decision Table, Random Tree, K*, SMO (an algorithm for training a support vector classifier), Simple Logistic, and Decision Tree. The reason why we choose them is that they are popular and also represent different categories of classification algorithms. Through those algorithms, we could find out which algorithm is the best fit for our short URL classification.

C. Classification Setup and Cross Validation

We know that click distributions from countries or referrers have no meaning if the user clicks is less than 2. Thus, our experiments will only focus on short URLs having that user clicks value is larger than or equals to 2 at least.

Several reasons have caused low clicks of URLs in our dataset. One reason for this is that the URL may be created recently compared with our dataset creation time so that our dataset is not able to collect more clicks. Also no interest from people and URL filtering of websites may cause low clicks as well. If spam URL attracts few clicks or is filtered by spam detection engine of website, that means this kind of URL is easy to distinguish or detect. We will not focus on this kind of URL in this paper.

In addition, we believe that the click traffic pattern of short URL spam appear more evidently as the increasing of user clicks. To prove that, we process dataset into 7 groups based on the value range of user clicks: ≥ 2 , ≥ 5, ≥ 10, ≥ 20, ≥ 30, ≥ 40, and ≥ 50. The reason for breaking-down the data into seven groups is as follows: first, group ≥ 2 serves the baseline group. After that, we want to increase the threshold by the same increase interval starting with the threshold 10. Due to that more than 30% short URLs are between threshold 2 and 10, the group ≥ 5 is added into the list. We try to show more accurate results by split the range between 2 and 10. For each group, we randomly choose the same amount of legitimate URLs as spam URLs to eliminate any prior probability influence.

We employed the machine learning classifiers previously mentioned using 10-fold cross-validation model. Cross validation is a technique for protecting against over-fitting in a predictive model. Specifically, the data is randomly divided into k groups and the classifier is re-estimated k times, holding back a different group each time. The overall accuracy of the classifier is the mean accuracy of the k classifiers tested.

VI. Evaluation

In this section, we first introduce the evaluation metrics for short URL classification. Then, we evaluate two major metrics that are the F-measure and accuracy of the classification framework based on the ground truth dataset.

A. Evaluation Metrics

We use two major metrics including the F-measure and accuracy to evaluate the performance of the classifiers. F-measure (also called F-score) is calculated based on precision and recall. Before introducing the details of precision and recall, we review in Table V the relationship between true positive, true negative, false positive, and false negative.

TABLE V. The relationship between true-positive, true-negative, false-positive, and false-positive.

Actual Label	Predicted Label	
	Positive	*Negative*
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Specifically, true positives are the number of instances that are correctly predicted as belonging to the positive class. True negatives are the number of instances that are correctly predicted as belonging to the negative class. False positives are the number of instances that are incorrectly predicted as belonging to the positive class. False negatives are the number of instances that are incorrectly predicted as belonging to the negative class. Based on these definitions, the formulas for Precision (P), Recall (R), F-measure (FM) and accuracy (A) are as follows [34]:

$$P = \frac{TP}{(TP + FP)},\ R = \frac{TP}{(TP + FN)} \tag{1}$$

$$FM = 2 \cdot \frac{P \cdot R}{P + R},\ A = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2}$$

Here, precision is obtained by dividing the number of the true positives by the sum of the true positives and false positives. Recall is obtained by dividing the number of the true positives by the sum of the true positives and false negatives. The goal of our experiment is to obtain high F-measure and accuracy values for better classification.

B. Evaluation Results

After performing all the classification experiments on seven groups (≥ 2 , ≥ 5, ≥ 10, ≥ 20, ≥ 30, ≥ 40, and ≥ 50) in our dataset, we present the results of cross validation in Figure 4. The results show that Random Tree, Random Forest, and K star algorithms outperform other four algorithms including Decision Tree, Decision Table, Simple Logistic, and SVM algorithms in terms of accuracy, F-measure and false positive rate. Especially, Random Tree algorithm performs the best among all seven algorithms. Meanwhile, as the number of user clicks increases, the performance of Random Tree, Random Forest, and K star classifiers has improved.

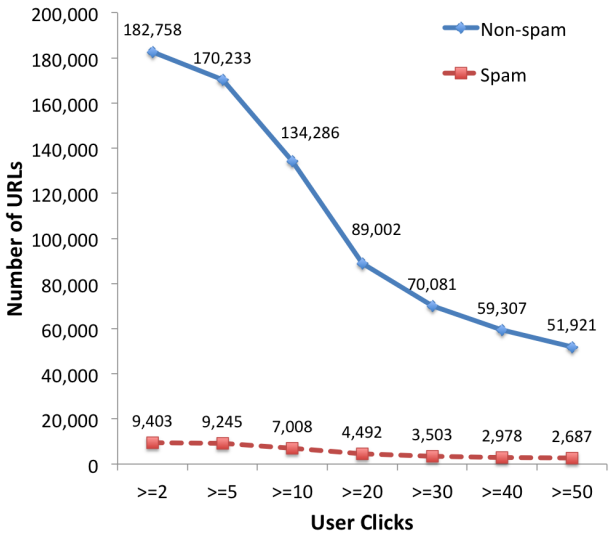
We also observe that the performance of some algorithms such as SVM has sharp drop in some groups like ≥ 20 and ≥ 30 groups. One possible reason is that the number of spam URLs has decreased a lot as we increase the threshold, which may result in overfitting in the classification, especially when the classifier is sensitive to the size of training dataset. However, other classifiers such as Random Tree and K star are stable as the increase of the threshold.

For the dataset in which user clicks number is larger than or equals to 2, we list four metrics for evaluating the classification performance sorted on accuracy in Table VI. It shows that the best classification performance is from the Random Tree algorithm but it still has high FP rate.

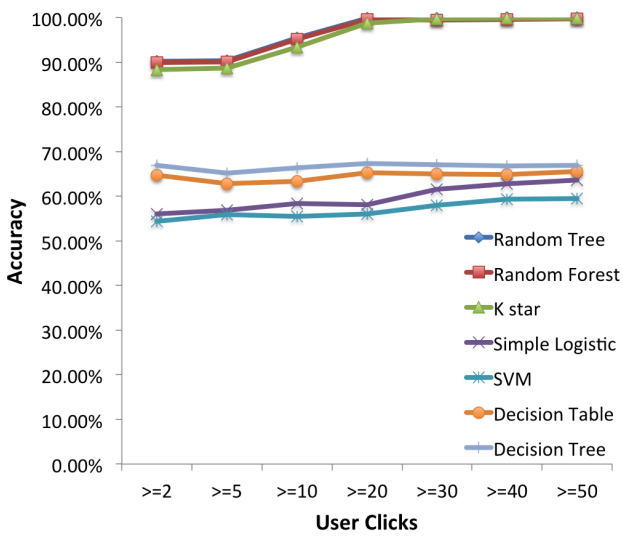
TABLE VI. Results of classification for short URL spam detection based on click traffic features

Algorithm	*TP*	*FP*	*F-measure*	*Acc.*
Random Tree	0.959	0.143	0.913	90.81%
Random Forest	0.946	0.134	0.910	90.6%
KStar	0.949	0.171	0.895	88.88%
Decision Tree	0.806	0.372	0.740	71.69%
Decision Table	0.622	0.342	0.634	64.03%
Simple Logistic	0.657	0.528	0.601	56.43%
SVM	0.886	0.807	0.658	53.93%

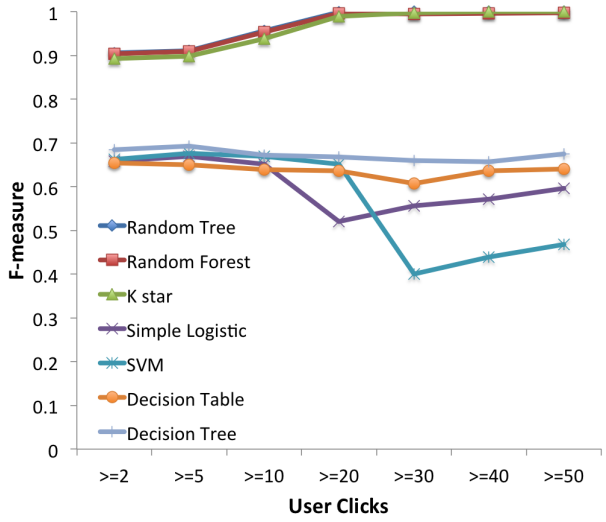
By investigating the errors, we attribute them to the following possible reasons:



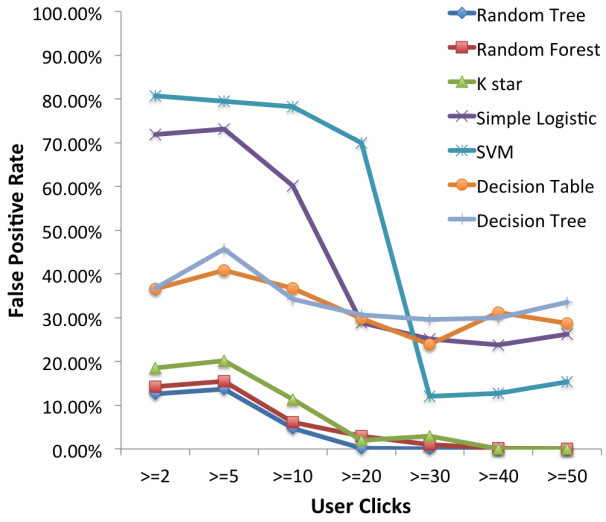
(a) Spam and Non-spam URLs in Datasets



(b) Accuracy



(c) F-measure



(d) False Positive Rate

Fig. 4. Experimental Results of Cross Validation

- Lack of features: only 9 click traffic features are used in classification, and
- Some mislabeled short URLs: Spammers might try to appear legitimate by mixing in spam and legitimate URLs in their posts. This might be as a result of previous spam fighting efforts by Twitter and following evolution of spammers.

In future work, we plan to investigate additional features for short URL classification and confirm labeling of a small sample of tweet URLs manually. And we will discuss this in more details later.

In classification, some features play a more important role than others during classification. Thus, by using information gain in feature ranking, we listed all sorted features in the decreasing order of information gain value (shown in Table VII). It evaluates the discrimination weight each feature has. The top

TABLE VII. RANKED FEATURES BASED ON INFORMATION GAIN

Feature Name	*Information Gain*
User clicks	0.0392
Standard deviation among country click sources	0.0387
Standard deviation among referrer click sources	0.0364
Mean among referrer click sources	0.0356
Global clicks	0.0308
Mean among country click sources	0.0289
Number of referrers	0.0219
Number of countries	0.019
Ratio between user clicks and global clicks	0.0174

3 discriminative features are user clicks, standard deviation among country click sources, and standard deviation among referrer click sources. It indicates there exists differences between spam and non-spam URLs in terms of distributions

of country click sources and referrer click sources. Number of referrers, number of countries, and ratio between user clicks and global clicks are the last 3 features. It implies that they did not help a lot in short URL spam classification.

By further investigating the correlation from those features to spam URLs, we looked into the value distributions of features in spam URLs and legitimate URLs. We found out that with the same mean among referrer click sources, there are more short URL spam when short URLs have low standard deviation among referrer click sources. It implies that spammers spread spam on a lot of referrers and the click traffic from those referrers show not much difference. Moreover, this phenomenon becomes more evident when the short URL obtains high click traffic.

C. Discussion

Although our classification shows good results in applying click traffic analysis on short URL spam detection, there are still many limitations and challenges we need to face.

1) Limitations: We have two major limitations in our classification with respect to two aspects: dataset collection and data labeling.

First, our click traffic dataset is based on APIs provided by URL shortening service providers. Thus, data collection is limited by those APIs. If service providers block the API access or modify their APIs, our data collection will need to modify accordingly. In addition, it is also hard for us to obtain other kinds of click traffic features outside the APIs such as daily clicks from country click sources and referrer click sources. Given daily click traffic features, we will be able to use them in classification and explore more special patterns in click traffic of spam URLs.

Second, we have used several public blacklists such as Google safe browsing, SURBL and URIBL in data labeling which provide strong validation of ground truth spam labels. But it is still possible that some URL spam in the dataset are mislabeled. One possible reason is that those blacklists keep updating and also removing old items based on their own policies. Thus, some URLs supposed to be spam URLs may be labeled as legitimate instead as the blacklists removed them from the list. Additionally, those mislabeling URLs exert more influences on the results of classification when the size of training dataset is small. Therefore, we need to obtain more validation resources for data labeling especially when our classification is deployed in real-time.

2) Challenges and Possible Countermeasures: In addition to limitations above, we are also facing several challenges in terms of performance and effectiveness in practice.

One challenge is that our classifiers work when the short URLs have click traffic (at least 2 clicks for each short URL). For those short URLs with less than 2 clicks, we ignored them since either no one is interested in the content or people recognize it as spam easily based on content. But only using our algorithm may be not enough for preventing spamming activities. Combining with other layers of spam detection may make a better result. For example, we analyze them based on content and user behaviors for those URLs with very few clicks. As they attract more clicks, we could combine our classification with behavior analysis like work done by Maia et al. [35] and content analysis [19].

Another challenge is that a spammer can setup a new URL shortener in several minutes to the same or another long URL spam after being detected. If the long URL is the same URL as the previous long URL or appears in the redirection chain to the previous long URL, we could store the previous long URL and the redirection chain to the previous long URL on our URL blacklist after we classified the short URL as spam URL. In such way, it will force spammers to create completely new domain to avoid detection. Our method could increase the cost of spamming activities of spammers at least.

Moreover, another challenge is that spammers could create click traffic for spam URL to confuse our classifier after they know our algorithm. We need to adopt methods for click fraud detection in our data pre-filtering process to eliminate the noise. This will be considered as future work.

VII. Related Work

Our work finds similarities to three main lines of work: social network spam, short URLs, and spamming on social networks via short URLs.

Social network spam has been investigated in several recent papers. Zhang et al. [36] proposed a method for detecting instances of automated Twitter accounts using the publicly available timestamp associated with each tweet. The work revealed that automated accounts exhibit distinct timing patterns that are detectable using statistical inference. Similarly, Castillo et al. [37] discussed how to assess the level of social media credibility of newsworthy topics from a given set of tweets, classifying those topics as credible or non-credible. Benevenuto et al. [38] have addressed the issue of detecting video spammers and promoters. Other recent works on spam detection in Twitter include Wang et al. [6] that proposed a classification approach to automatically identify suspicious users by (i) a directed social graph model that captures the follower and friend relationships in the network, and (ii) content-based and graph-based features extracted based on the spam policy of Twitter, and Benevenuto et al. [4] that introduced machine learning techniques to identify spammers based on number of followers, number of followees, and other social interactions such as the number of times the user was mentioned and the number of times the user was replied to.

The popularity of short URLs has immensely increased over the years due to micro-blogging platforms such as Twitter, where message space is limited by 140 characters. Antoniades et al. [13] has recently explored this emerging phenomenon and presented a characterization of short URLs. Their analysis was performed on a dataset collected by crawling Twitter for short URLs from Bitly and Owly URL shortening services. Specifically, their results showed that (i) the maximum access to short URLs come from emails and online social media, (ii) the click distribution of short URLs is approximately a log-normal curve, and (iii) a large percentage of short URLs are not ephemeral and 50% of short URLs live for more than three months.

Recently, several papers have investigated spamming on social networks via short URLs. Grier et al. [1] presented a

characterization of spam on Twitter. Their analysis showed that (i) Twitter spam is more effective than email spam with an overall clickthrough rate of 0.13%, (ii) blacklists are no optimal solution for fighting spam on Twitter as they are too slow at identifying new threats, and (iii) spammers use URL shortener services to obfuscate their links in tweets, negating any potential gains even if blacklist delays were reduced. Chhabra et al. [10] analyzed phishing attacks on Twitter using URL shorteners. Phishing is one form of spam, where the goal is to steal personal information from users for fraudulent purposes. This work concluded that (i) phishers use URL shorteners to hide their identity, (ii) online social media brands such as Twitter are targeted by phishers more than traditional brands such as eBay, and (iii) phishing URLs which are referred from Twitter are more likely to attract victims. Klien et al. [39] studied usage logs of a URL shortener service that had been operated by the authors for more than a year. Their results showed that (i) different countries differ significantly with regard to the usage of their service, (ii) around 80% of URLs shortened by their service lead to spam-related content, and (iii) spamming attacks via short URLs cross national borders. Maggi et al [14] measured two years of short URLs and provided some countermeasures but it did not offer efficient short URL spam detection approach.

Our work differs from the aforementioned research along three dimensions: (i) we analyze a comprehensive dataset containing over $600,000$ short URLs; (ii) we consider spam in general and do not restrict the analysis to a specific form of spam such as phishing, and most importantly; (iii) we attempt to classify short URLs as to whether they lead to spam or not using their click traffic information.

VIII. Conclusion

We conducted the first large-scale experimental study of short URLs through creator and click source analysis on the Bitly dataset - a collection of 641,423 short URLs. We first analyzed the creators of the short URLs and determined that the legitimate creators in Bitly generate short URL spam as well. As future work, we plan to uncover spam creators after short URL classification. We then examined the clicks to the short URLs and found that the majority of the clicks are from direct sources such as email clients and that the spammers utilize popular websites such as Facebook to attract more attention. We finally performed classification of short URL spam based on click traffic and analyzed performance change of classifiers as the increase of user clicks. Random Tree, Random Forest, and K start algorithms outperform other algorithms. Of them, the Random Tree algorithm achieved the best performance with an accuracy of 90.81% and an F1-measure value of 0.913. We believe some of the classification errors might have been caused by the lack of features and some mislabeling in the dataset.

Our analysis and classification work can be considered as a new approach to classification in the ongoing battle of short URL spam detection. An interesting direction for future research involves combing our click traffic analysis process with these existing analysis techniques to create a multi-layered defense against short URL spam.

Acknowledgements

This research has been partially funded by National Science Foundation by CNS/SAVI (1250260), IUCRC/FRP (1127904) , CISE/CNS (1138666), RAPID (1138666), CISE/CRI (0855180), NetSE (0905493) programs, and gifts, grants, or contracts from DARPA/I2O, Singapore Government, Fujitsu Labs, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

References

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

[2] 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 2011 IEEE Symposium on Security and Privacy*, ser. SP '11, 2011.

[3] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th Annual Computer Security Applications Conference*, ser. ACSAC '10, 2010, pp. 1–9.

[4] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Proceedings of the Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, 2010.

[5] F. Benevenuto, H. Haddadi, and K. Gummadi, "The World of Connections and Information Flow in Twitter," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, no. 4, pp. 991–998, Jul. 2012.

[6] A. H. Wang, "Don't follow me: Spam detection in twitter," in *Proceedings of the International Conference on Security and Cryptography*, 2010.

[7] S. Yardi, D. M. Romero, G. Schoenebeck, and D. Boyd, "Detecting spam in a twitter network," *First Monday*, vol. 15, no. 1, 2010.

[8] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *Proceedings of the Internet Measurement Conference*, 2010.

[9] A. Neumann, J. Barnickel, and U. Meyer, "Security and privacy implications of url shortening services," in *WEB 2.0 Security & Privacy Workshop(W2SP)*, 2011.

[10] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru, "Phi.sh/$ocial: the phishing landscape through short urls," in *Proceedings of the Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, 2011.

[11] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, "Understanding and combating link farming in the twitter social network," in *Proceedings of the 21st international conference on World Wide Web*, ser. WWW '12, 2012, pp. 61–70.

[12] K. C. T, H. Ponnapalli, D. Herts, and J. Pablo, "Analysis and Detection of Modern Spam Techniques on Social Networking Sites," in *2012 Third International Conference on Services in Emerging Markets*. IEEE, Dec. 2012, pp. 147–152.

[13] D. Antoniades, I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ioannidis, E. P. Markatos, and T. Karagiannis, "we.b: the web of short urls," in *Proceedings of the International Conference on World Wide Web*, 2011.

[14] F. Maggi, A. Frossi, S. Zanero, G. Stringhini, B. Stone-Gross, C. Kruegel, and G. Vigna, "Two years of short urls internet measurement: security threats and countermeasures," in *Proceedings of the 22nd international conference on World Wide Web*, ser. WWW '13. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2013, pp. 861–872.

[15] X. Jin, C. X. Lin, J. Luo, and J. Han, "Socialspamguard: A data mining-based spam detection system for social media networks," in *Proceedings of the International Conference on Very Large Data Bases*, 2011.

[16] D. Wang, D. Irani, and C. Pu, "A social-spam detection framework," in *Proceedings of the Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, 2011.

[17] Google Safe Browsing API, https://developers.google.com/safe-browsing/, 2013, accessed on August. 1, 2013.

[18] D. Fetterly, M. Manasse, and M. Najork, "Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages," in *Proceedings of the International Workshop on the Web and Databases*, 2004.

[19] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in *Proceedings of the International World Wide Web Conference*, 2004.

[20] A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using traffic feature distributions," in *Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, ser. SIGCOMM '05, 2005.

[21] P. H. Las-Casas, D. Guedes, J. M. Almeida, A. Ziviani, and H. T. Marques-Neto, "Spades: Detecting spammers at the source network," *Computer Networks*, vol. 57, no. 2, pp. 526 – 539, 2013.

[22] Twitter Developers, https://dev.twitter.com/, 2013, accessed on August. 1, 2013.

[23] WWW::Shorten::Bitly - Interface in Perl, http://search.cpan.org/~pjain/WWW-Shorten-Bitly-1.17/lib/WWW/Shorten/Bitly.pm, 2013, accessed on August. 1, 2013.

[24] About Twitter's Link Service, http://support.twitter.com//entries/109623, 2013, accessed on August. 1, 2013.

[25] S. Webb, J. Caverlee, and C. Pu, "Introducing the webb spam corpus: Using email spam to identify web spam automatically," in *Proceedings of the Conference on Email and Anti-Spam*, 2006.

[26] McAfee SiteAdvisor, http://www.siteadvisor.com/, 2013, accessed on August. 1, 2013.

[27] URIBL, http://www.uribl.com/, 2013, accessed on August. 1, 2013.

[28] SURBL, http://www.surbl.org/, 2013, accessed on August. 1, 2013.

[29] The Spamhaus Project, http://www.spamhaus.org/, 2013, accessed on August. 1, 2013.

[30] Bitly Enterprise, http://www.enterprise.bitly.com/, 2013, accessed on August. 1, 2013.

[31] H. Kwak, C. Lee, H. Park, and S. B. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the International Conference on World Wide Web*, 2010.

[32] C. E. Landwehr, D. Boneh, J. C. Mitchell, S. M. Bellovin, S. Landau, and M. E. Lesk, "Privacy and cybersecurity: The next 100 years," *Proceedings of the IEEE*, vol. 100, no. Centennial-Issue, pp. 1659–1673, 2012.

[33] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[34] N. Chinchor, "Muc-4 evaluation metrics," in *The Fourth Message Understanding Conference*, 1992.

[35] M. Maia, J. Almeida, and V. Almeida, "Identifying user behavior in online social networks," in *the 1st workshop on Social network*. New York, New York, USA: ACM Press, 2008, pp. 1–6.

[36] C. M. Zhang and V. Paxson, "Detecting and analyzing automated activity on twitter," in *Proceedings of the Passive and Active Measurement Conference*, 2011.

[37] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the International Conference on World Wide Web*, 2011.

[38] F. Benevenuto, T. Rodrigues, A. Veloso, J. M. Almeida, M. A. Gonçalves, and V. A. F. Almeida, "Practical detection of spammers and content promoters in online video sharing systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 42, no. 3, pp. 688–701, 2012.

[39] F. Klien and M. Strohmaier, "Short links under attack: Geographical analysis of spam in a url shortener network," in *Proceedings of the ACM Conference on Hypertext and Hypermedia*, 2012.