AN IN-DEPTH ANALYSIS OF ABUSE ON TWITTER

Jonathan Oliver
Trend Micro, Australia

Paul Pajares
Trend Micro, Philippines

Christopher Ke, Chao Chen & Yang Xiang
Deakin University, Australia

Email {jon_oliver, paul_pajares}@trendmicro.com, {christopher.ke, zvm, yang.xiang}@deakin.edu.au

ABSTRACT

In this paper, we examine *Twitter* in depth, including a study of 500,000,000 Tweets collected over a two-week period in order to analyse how the micro-blogging site is abused.

Most *Twitter* abuse takes the form of Tweets containing links to malicious websites. These websites take many forms, including spam sites, scam sites that are involved in compromising more *Twitter* accounts, phishing sites, and sites hosting malware or offering cracked versions of software. Many of the malicious Tweets are sent from legitimate accounts that have been compromised, causing a range of problems for their owners.

The scale of the threat is significant. Previous research, notably [1], has indicated that the use of URL blacklists is ineffective in detecting *Twitter* threats. However, our research shows otherwise – approximately 5% of all Tweets with links had malicious and/or spammy content.

We also applied graph algorithms to the *Twitter* data and were able to find various clusters of interrelated websites and accounts. We were able to identify specific Tweet spam campaigns as well as the groups carrying out these campaigns.

The data from this analysis leads us to conclude that blacklisting, in conjunction with other analytical tools, is an effective tool for identifying malicious Tweets.

1. INTRODUCTION

Researchers from *Trend Micro* and Deakin University worked together to investigate the *Twitter* threat landscape. This paper features a comprehensive study that lasted for two weeks from 25 September to 9 October 2013, including further analysis of some of the threats we discovered over the given period. The study revealed a significant level of abuse of *Twitter*, including spamming, phishing, and sharing of links that led to malicious and potentially illegal websites. The majority of the malicious messages we observed were sent from compromised accounts, many of which have subsequently been suspended by *Twitter*.

A 2010 study [1] examined 400 million public Tweets and 25 million URLs. The authors identified two million URLs (8%) that pointed to spamming, malware-download, scamming and phishing websites, leading them to conclude (a) that blacklists were ineffective, as these only protected a minority of users, and (b) that the use of URL shorteners made the task of identifying malicious links very difficult.

This research paper begins by giving a brief overview of the types of *Twitter* abuse we discovered within our study period. It then provides a summary of the data we collected to learn more about the abuse. Given the data, we examined a range of issues, including: (a) the use of blacklists to detect *Twitter* spam, (b) the coordinated nature of certain *Twitter* spam outbreaks, (c) the timing of spam outbreaks, and (d) details related to particular *Twitter* scams. In Section 4, we propose an approach for analysing *Twitter* spam outbreaks which is very useful in augmenting blacklists for the detection of *Twitter* spam.

2. OVERVIEW OF THE ABUSE ON TWITTER

This Section provides a brief overview of the *Twitter* threats we found. It also provides examples of the most active threat types, including: traditional spam similar to email spam, searchable spam (which differed from email spam), phishing messages, and suspended and compromised accounts.

2.1 Traditional spam

The following are some of the features of traditional *Twitter* spam:

- The Tweets typically promoted weight-loss drugs, designer sunglasses and bags, etc., very much like email spam.
- Unrelated, but often-trending hash tags were used to increase Tweet distribution and to encourage more people to click the links.
- The Tweets included misspelled words, sometimes substituting numbers for letters, which was typical of email spam 10 years ago.
- In some cases, URL shorteners were used to make it more difficult for security analysts to identify which Tweets point to spam websites.

2.2 Searchable spam

Figure 1 shows examples of searchable *Twitter* spam.

The following are some of the features of searchable spam:

- The messages typically promote free access to copyrighted and licensed materials or offer gadget

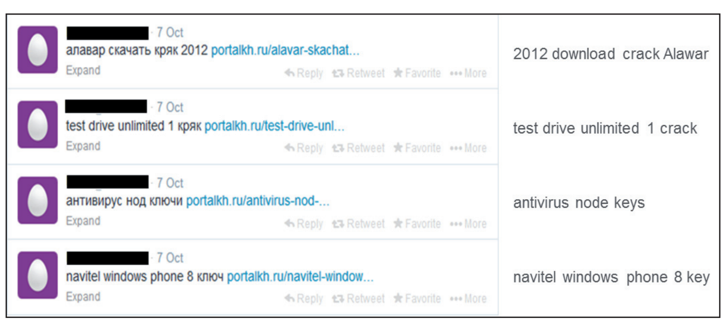

Figure 1: Examples of searchable spam (translations on the right).

knock-offs. For example: solutions to homework and exam cheat sheets, free movie downloads, cracked versions of software, and computer, printer and mobile device knock-offs.

- Hash tags are not used, or are only used sparingly.
- Many such Tweets are written in Russian.
- Several domains are used, many of which are hosted in Russia and in the Ukraine.

Our analysis of searchable spam revealed that the probability of *Twitter* suspending an account involved with a searchable spam incident was significantly lower than if it was involved in sending out traditional *Twitter* spam or other malicious messages. In addition, we found that 50% of those who clicked the links in searchable spam written in Russian were from non-Russian-speaking countries such as the United States and Japan (see Section 7). This type of spam typically remains on *Twitter* after transmission and can easily be searched for. For example, Group A, described in Section 5, consists of over 7.8 million searchable spam messages. Approximately 90% of these remain accessible on *Twitter* at the time of writing this paper.

We conclude that searchable spam attempts to avoid irritating users so that it will not be reported via the 'Abuse' button that *Twitter* has made available. Searchable spam covers a wide range of content, which some users might be motivated to look for using *Twitter*'s Search function. They might even be willing to use automated translation tools to understand the content of such spam.

2.3 Twitter phishing

We examined a long-running phishing scam [2] that exploits certain *Twitter* features. The scam starts with a compromised user sending messages to friends (using the @ syntax on *Twitter*). The messages ask them to click a shortened URL – clicking the link starts a redirection chain that ends at a phishing page that tells the user their session has timed out and that they need to log in again. In the course of our research, we attempted to estimate the scale of this problem, which we discuss in Section 8.

2.4 Suspended and compromised accounts

While carrying out our research, we followed some of the accounts that had been involved in spamming. We attempted to access them in December 2013 (two months after the period of data collection). We found that *Twitter* had suspended tens of thousands of accounts involved in spamming and in other malicious activities. Many of these accounts appeared to have been created specially for this purpose – the accounts were created, and then immediately started sending spam. In some cases, genuine account owners had identified the problem and taken corrective actions to restore their accounts. However, this was significantly rarer than account suspension. (We do not have statistics on this because it was difficult to establish when compromises occurred; we only have anecdotal evidence of their occurrence.)

3. RESEARCH SCOPE AND METHODOLOGY

We collected as many Tweets with embedded URLs as possible within the two-week period from 25 September to 9 October 2013. We restricted the Tweets we examined to those with embedded URLs. While it is possible to use *Twitter* to send spam and other messages without URLs, the majority of the spam and other malicious messages we found on *Twitter* contained embedded URLs. Among the thousands of spam messages that humans inspected in the course of our research, we only found a handful of Tweets without URLs that could be considered abusive or harmful.

We categorize Tweets that contain malicious URLs as 'malicious Tweets'. The data we collected is shown in Table 1. We gathered a total of 573.5 million Tweets containing URLs and identified

Day/date	Number of Tweets with URLs	Number of malicious Tweets	Percentage of malicious Tweets
Wednesday 09/25/2013	39,257,353	2,292,488	5.8%
Thursday 09/26/2013	47,252,411	3,190,600	6.8%
Friday 09/27/2013	49,465,975	3,947,515	8.0%
Saturday 09/28/2013	37,806,326	2,018,935	5.3%
Sunday 09/29/2013	-	-	-
Monday 09/30/2013	-	-	-
Tuesday 10/1/2013	48,778,630	2,511,489	5.1%
Wednesday 10/2/2013	51,728,355	3,739,597	7.2%
Thursday 10/3/2013	51,638,205	3,932,186	7.6%
Friday 10/4/2013	49,230,861	3,398,526	6.9%
Saturday 10/5/2013	44165664	2293539	5.2%
Sunday 10/6/2013	45,089,730	2,006,447	4.4%
Monday 10/7/2013	50,457,403	2,305,794	4.6%
Tuesday 10/8/2013	42,031,232	1,152,119	2.7%
Wednesday 10/9/2013	16,612,318	538,133	3.2%
TOTAL	**573,514,463**	**33,327,368**	**5.8%**

Table 1: Data collected.

33.3 million malicious Tweets, which accounted for approximately 5.8% of all of the Tweets with URLs[1]. We used two methods to identify malicious Tweets. The first involved the use of the *Trend Micro* Web Reputation Technology [3], which uses a blacklist. The second method involved identifying groups of malicious Tweets using the clustering algorithm described in Section 4. Note that we experienced a disruption in our data-collection process on 29 and 30 September 2013, which accounted for data loss during said period.

4. CLUSTERING ALGORITHM TO IDENTIFY MALICIOUS TWEETS

One of our research goals was to obtain a high-level understanding of the various types of spam and scams on *Twitter*. We determined that one approach to achieving this aim would be to cluster malicious Tweets into groups. Forming clusters of malicious Tweets would be successful if we could explain adequately why Tweets in a group are considered similar to one another, and why they are considered malicious.

Several possible variables could be extracted from Tweets, including: content, embedded URLs, hash tags and sender data, including frequency. It would prove very useful if it were possible to group *Twitter* spam into distinct outbreaks rather than try to understand a huge mass of data. Traditional approaches for doing this include grouping spam Tweets that have similar content or applying machine-learning approaches. Applying machine-learning approaches involves extracting numerical or categorical variables from Tweets and users (e.g. how often they send messages, dramatic changes in their behaviour, etc.) and applying a statistical or machine-learning approach to the data (e.g. SVMs or Nearest Neighbor).

We took another approach. Our proposal for identifying certain classes of high-volume spam was to create a graph consisting of senders and domains in URLs and to identify bipartite cliques [4] in this graph. Such graphical approaches to identifying cliques in data have previously been applied to computer security problems [5]. To do this, we constructed a graph where the *Twitter* users are nodes on the left-hand side of the graph while the domains in links are nodes on the right-hand side. For each Tweet from User U that contains a link with Domain D, we include an arc in the graph from User U to Domain D. Some spammers use applications that employ a round-robin approach for sending spam. Given a number of sending accounts and destinations for URLs in the Tweets, the use of a round-robin approach maximizes the number of spam messages while minimizing the effects of (i) having their accounts suspended and (ii) blacklists blocking their spam. When the graphical approach described above is used, a set of users involved in a round-robin approach will generate a bipartite clique in the graph. Hence, bipartite cliques in such a graph are very suspicious – the probability of real users behaving this way in the normal course of events is extraordinarily small. There are scalable approaches for using map-reduce [6, 7] to identify cliques in large data sets.

Figure 2 provides an example of a bipartite clique found in the data, which consists of 727 users who sent Tweets containing links to 11 domains; all of the users in the clique sent Tweets containing links to all of the domains in the clique.

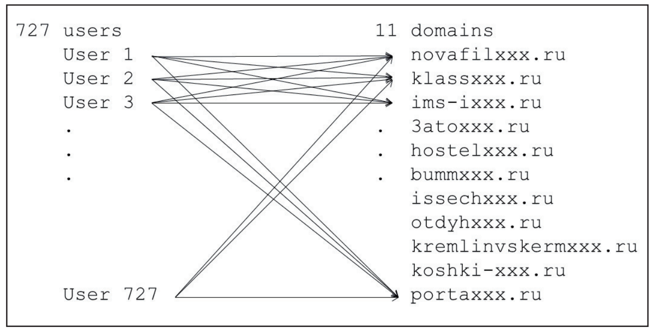

Figure 2: Sample bipartite clique.

This approach is well suited to understanding certain types of *Twitter* spamming behaviours, but unsuited to others. For example, it is not suitable for analysing the *Twitter* follower scam described in Section 6, since that did not use a round-robin approach for sending messages. The *Twitter* follower scam was confirmed as malicious by installing the app and monitoring its behaviour.

Other malicious behaviour was identified by following the links through to the final website and confirming that the website was malicious.

5. HIGH-LEVEL PERSPECTIVE

We applied the clique algorithm described in Section 4 [6] to the *Twitter* data we collected. The algorithm identified 16 cliques, each of which accounted for 1% or more of the *Twitter* spam. Table 2 describes each of the cliques generated. In addition, Group G was a *Twitter* follower spam group, which accounted for 2.5% of the *Twitter* spam.

The columns in Table 2 are defined as follows:

- The 'Description' column describes the content of the Tweets.
- The 'Percentage of malicious Tweets' column gives the percentage of Tweets out of the total 28 million in the group.
- The 'Senders' column shows the number of confirmed senders in a clique. As such, a confirmed sender should have sent Tweets to all of the domains in a clique. For example, 797 senders sent at least 24 messages with links going to all of the 24 domains in Group A. The number of senders in Group G is simply the number of senders who sent Tweets with URLs that led to a *Twitter* follower scam website. In this case, there was no convenient confirmation step to separate legitimate users who had re-Tweeted spam from those whose accounts were under spammers' control.
- The 'Hash tags' column summarizes the use of hash tags in spam that belong to the group.
- The 'Domains' column lists the number of domains. Some groups used multiple hosts from the same domain. For example, Group H had five separate domains and used 10 distinct hosts on each of them.

[1] The authors understand that the two-week study period was during a period of spam activity that was significantly higher than the norm.

Description	Percentage of malicious Tweets	Number of senders	Hash tags	Number of domains	Percentage of suspended accounts
A. Education spam, etc.	27.28%	797	None	24	10.3%
B. Cracked software and game spam	8.11%	578	None	20	31.5%
C. Education spam	6.26%	539	None	20	19.7%
D. Cracked software spam	6.19%	9,509	Limited[1]	21	12.0%
E. Cracked software spam	4.39%	727	None	11	11.6%
F. Printer/mobile spam	3.72%	12,275	Low	3	89.1%
G. *Twitter* follower spam	2.54%	59,205	Yes	1	2.1%
H. Video/mobile/cracked software/game spam	2.23%	8,987	Low	50	95.2%
I. Game and computer spam	2.04%	608	None	19	97.9%
J. Education spam, etc.	1.99%	284	None	14	47.9%
K. Shirt spam	1.91%	1,699	None	5	74.7%
L. Game, mobile, and printer spam	1.81%	1,197	None	18	98.8%
M. Computer/printer spam	1.77%	26,603	Low	60	42.3%
N. Game/hardware spam	1.53%	2,514	Yes[2]	70	90.0%
O. Computer game/mobile device spam	1.41%	1,491	None	73	94.7%
P. Credit and education spam	1.08%	8,541	None	32	72.5%
Q. Cracked software and game spam	1.02%	9,066	None	4	98.6%
Other spam	24.74%	N/A	N/A	N/A	N/A

[1] 12.5% of the users from this group included the hash tag '#fgsdfg,' which has been used in subsequent spam outbreaks. The most recent outbreak (at the time of writing this paper) was seen on 8 January 2014.

[2] The common hash tags from this group included '#www,' '#Windows,' 'GTA (Grand Theft Auto)' and '#Samsung.'

Table 2: High-level perspective.

- The 'Percentage of suspended accounts' column shows the percentage of accounts that had been suspended when we checked their status in December 2013 – two months after the study period.

We note the following from Table 2:

- The 17 groups listed account for 75% of the *Twitter* spam we identified.
- It is highly likely that there were other types of abuse and spam that we were not able to identify in the study.
- *Twitter* responds very effectively to some spam outbreaks. For example, it identified and suspended over 95% of the accounts in Groups H, I, L and Q. Other spamming behaviours were not detected. For example, in Group A, which accounted for over 27% of the spam we found, approximately 10% of the accounts were suspended.

6. DETAILS ON SPECIFIC OUTBREAKS

6.1 Russian-138 spam

Six of the groups described in Section 5 had a set of features in common. We coined the term 'Russian-138 spam' to describe *Twitter* spam with the following features:

- The Tweets were primarily written in Russian.
- Many of the domains in the Tweets were .ru domains.
- The URLs were followed by a date stamp.

For example, a Tweet with the URL http://xxxxxx.ru/angliyskiy-fizik-moss-t-1380765135.html was sent on 5 October 2013. '1380765135' appears to be a time stamp that translates to 'Thursday 3 October, 01:52:15 2013 UTC', two days before the Tweet was sent.

The six groups that were characterized as Russian-138 spam were Groups A, B, C, E, I and J. Figure 3 shows the number of Tweets per hour in each of the groups monitored within the study period.

Figure 3 highlights the spammy nature of the groups:

- The groups of spamming *Twitter* users are acting in a coordinated manner. They start and stop spamming at roughly the same time.
- In some situations, one group of users will stop spamming to a set of domains while at the same time another group will start spamming another set of domains. Examples of this include the following:
  (i) At 2013-10-04 11:00 UTC, Group A (blue) stopped spamming and Group C (yellow) started spamming.

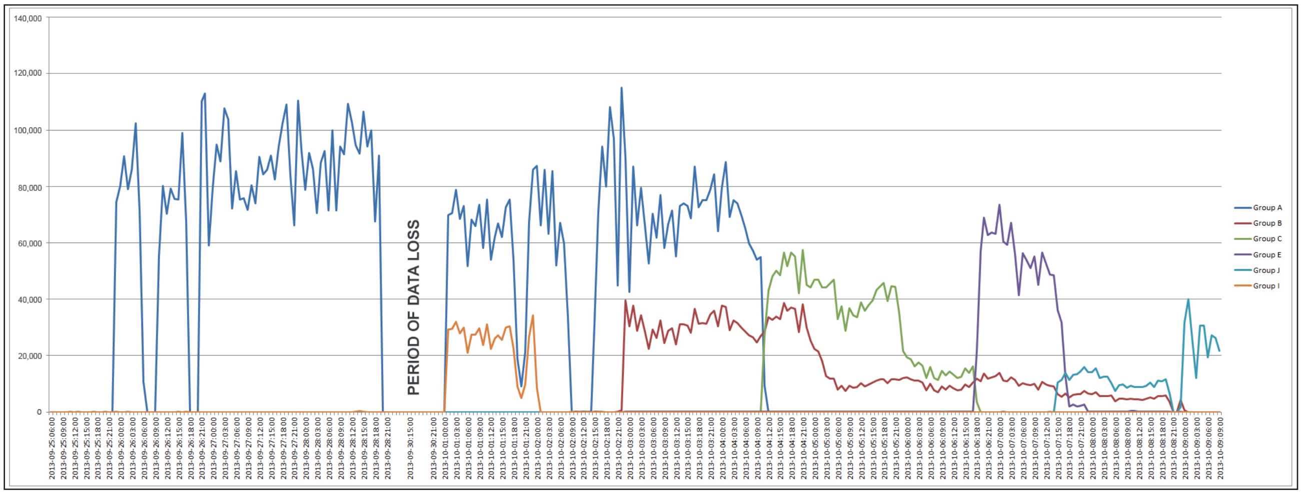

Figure 3: Number of Tweets per hour for the six Russian-138 spam groups.

Host details	Number of domains	Sample domains
IP address: 68.178.255.65, ns1.dershanelerkapatilmasin.com, ns2.dershanelerkapatilmasin.com **Country:** United States **ASN:** 26496	35	askfollow.com, askfollow.net, bestfollow.info, worldfollowers.info, etc.
IP address: 208.109.108.124, ns1.ip-68-178-255-209.secureserver.net, ns2.ip-68-178-255-209.secureserver.net **Country:** United States **ASN:** 26496	26	bestfollowers.org, biturlx.com, bulkfollowers.co, utf8more.info, etc.
IP address: 69.175.70.173, ns05.domaincontrol.com, ns06.domaincontrol.com **Country:** United States **ASN:** 32475	26	15c.info, azmh.info, cefpua.info, yigm.info, etc.
IP address: 172.70.175.69, 69.175.70.172, ns35.domaincontrol.com, ns36.domaincontrol.com **Country:** Namibia, United States **ASN:** 32475	5	followback.info, hitfollow.info, letgetmorefollowers.info, newfollow.info, plusfollower.info
IP address: 54.225.82.214, ns75.domaincontrol.com, ns76.domaincontrol.com **Country:** United States **ASN:** 14618	7	ferrastudios.com, followmania.co, followmania.com, unfollow.ferrastudios.com, etc.

Table 3: Summary of Twitter scam infrastructure and domains.

(ii) At 2013-10-06 18:00 UTC, Group C (yellow) stopped spamming and Group E (black) started spamming.

6.2 Twitter follower scams

In January 2014, we reported a *Twitter* follower scam [8] that used spam to entice users to install an app and authorize its access to their accounts. Once authorization was granted, a user's account would get more followers (i.e. other users of the app), become a follower of other users of the app, and possibly send out *Twitter* spam advertising the app. The IP addresses that host the scam are shown in Table 3. The majority of the victims were from the United States and Turkey. The premium service access prices were €5–10.

At the end of January 2014, we saw a spike in the number of users attempting to visit sites involved with scams, as shown in Figure 4. Hundreds of users attempted to access domains that contained instructions that, if followed, would cause their *Twitter* accounts to be compromised. Figure 4 also shows the

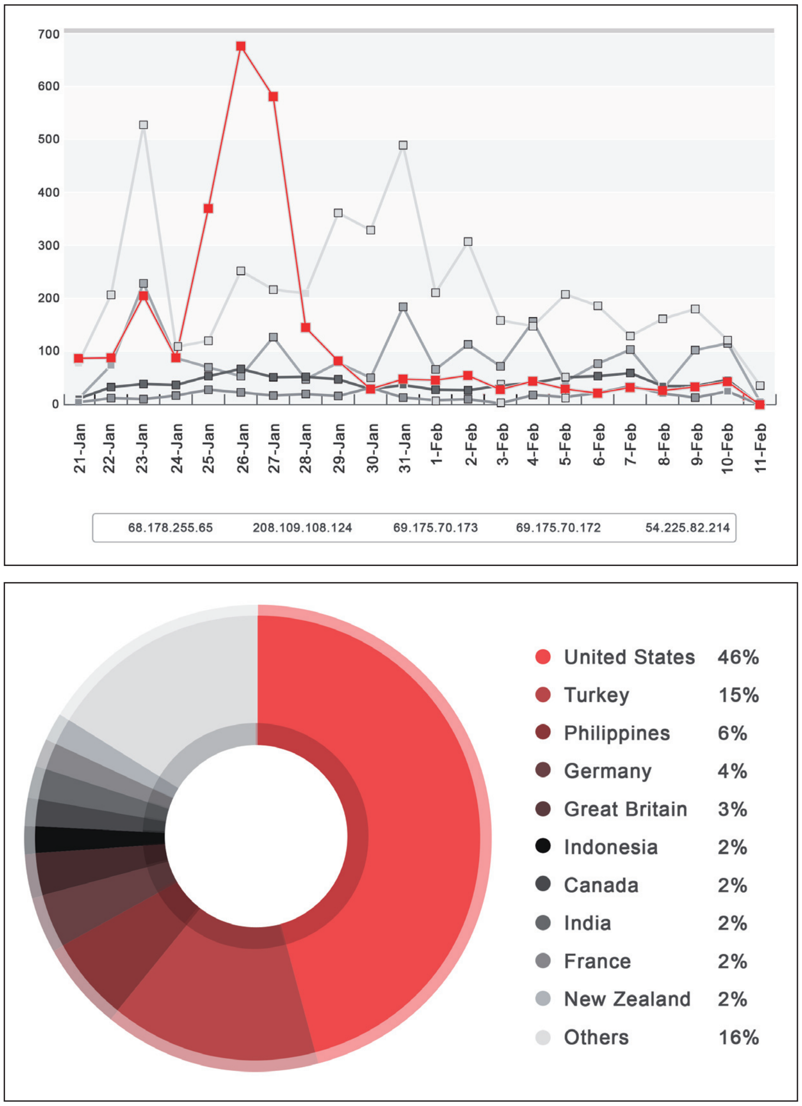

Figure 4: Impact of Twitter scams from January to February 2014.

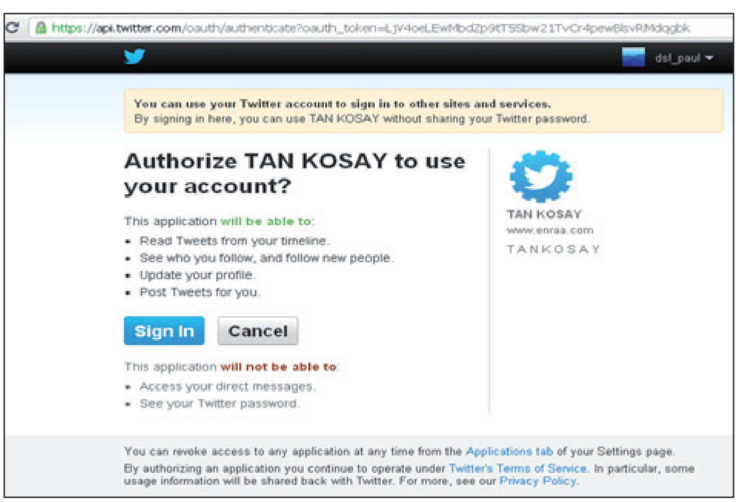

Figure 5: Authorizing Twitter-related apps.

distribution of users that were targeted by this scam, the majority of whom were from the United States. A significant number also came from Turkey, most likely because of the keyword 'takip' in some of the domains, which means 'follow up' in Turkish. The content of most of the web pages was written in English so users from the US could be their primary targets.

Users must be cautious of allowing third-party apps access to their *Twitter* accounts (see Figure 5). If they have been victimized by the scam above, they should revoke the malicious app's access rights through their settings.

7. IMPACT ANALYSIS OF CLICK-THROUGH DATA

Previous studies on email spam [9–11] have found that click-through and conversion rates vary considerably. The estimated click-through rates (i.e. the number of people who click a link in an email and thus arrive at a particular website) ranged from 0.003% to 0.02% [9, 10]. The 2010 study [1] on *Twitter* spam estimated the click-through rate at 0.13%, which suggests that the click-through rate for *Twitter* spam was two orders of magnitude higher than for email spam.

The *Trend Micro* Web Reputation Technology [3] has a component that allows users to obtain malicious anonymized feedback if they wish to. We examined the feedback data to determine which malicious URLs embedded in Tweets had been clicked. Without access to the platform's backend infrastructure, it was difficult to determine the absolute *Twitter* spam click-through rate. However, we were able to sensibly compare the relative effectiveness of malicious campaigns and determined that there was great variability between campaigns.

We classified the groups and domains we analysed in Section 5 into the following categories:

- **Malware:** Tweets with embedded links that led to malware-distribution websites.
- **Traditional phishing:** Tweets with embedded links that led to phishing websites.
- **Twitter-specific scam:** Tweets that led to the *Twitter* follower scam described in Section 6.
- **Spam:** Tweets that were sent by groups or domains involved in spam distribution. We split this category into four subcategories because the different spam flavours had distinct characteristics.

  The subcategories are:

  (i) Traditional spam

  (ii) Spam with shortened URLs

(iii) Russian spam, including the most prolific type, Russian-138 spam, described in Section 6

(iv) Spam related to a viral Japanese campaign.

There were enormous variations in the effectiveness of the different approaches to *Twitter* spamming. For example, the viral Japanese campaign was approximately 5,000 times more effective than the Russian spam campaign.

Abuse category	Clicks per Tweet
Viral Japanese spam campaign	0.26862
Malware	0.03065
Traditional phishing	0.00959
Spam with shortened URLs	0.00388
Spam	0.00239
Twitter-specific scam	0.00112
Russian spam	0.00005

Table 4: Clicks per Tweet.

7.1 Viral Japanese spam campaign

The viral Japanese spam campaign continued until February 2014. The vast majority (99%+) of users that were victimized were Japanese.

7.2 Malware Tweets

While conducting the study, we witnessed an outbreak of Arabic Tweets with embedded links that led to malware-laden websites. The majority of the affected users were from Saudi Arabia, Egypt and Sudan, followed by the United States (see Figure 6).

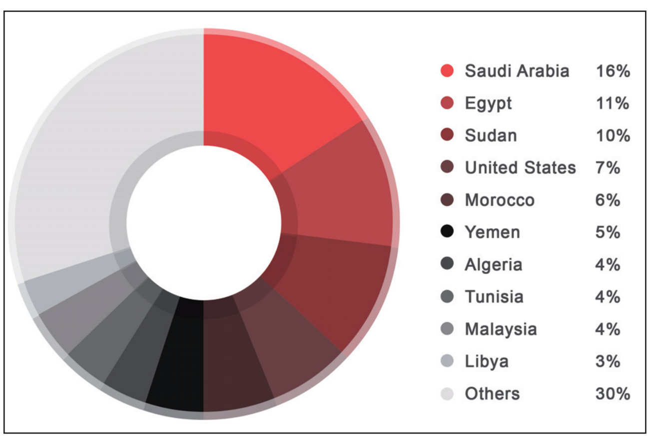

Figure 6: Distribution of clicks that led to malware-laden websites.

6.3 Traditional phishing Tweets

Traditional phishing Tweets are similar to phishing emails. The Tweets attempt to convince users that they came from legitimate users. As shown in Figure 7, the phishing Tweets we studied primarily targeted users in the United States.

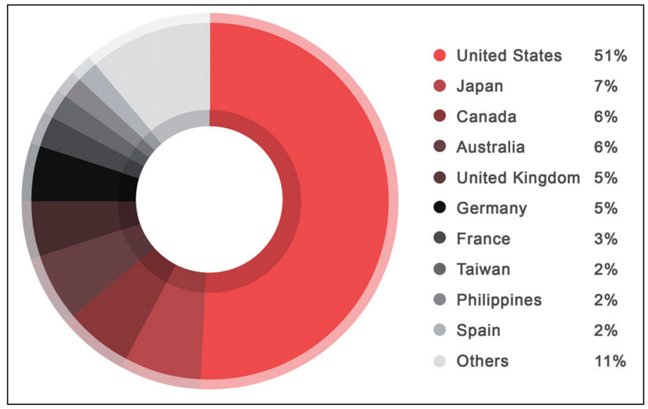

Figure 7: Distribution of clicks that led to phishing websites.

7.4 Spam with shortened URLs

A range of URL shorteners and proxy-avoidance domains were also used to obscure links in Tweets. This issue was discussed at length in the 2010 study on *Twitter* spam [1]. Within our study period, apart from the commonly abused bit.ly shortener, we also saw URL shorteners such as 17q.org, bitlyjmp.com, kisalink.tk, lima.pp.ua, qwapo.es, redir.ec, shortredirect.us and shortn.me used in malicious Tweets. The distribution in Figure 8 reflects the use of region-specific URL shorteners such as kisalink.tk and qwapo.es in some outbreaks.

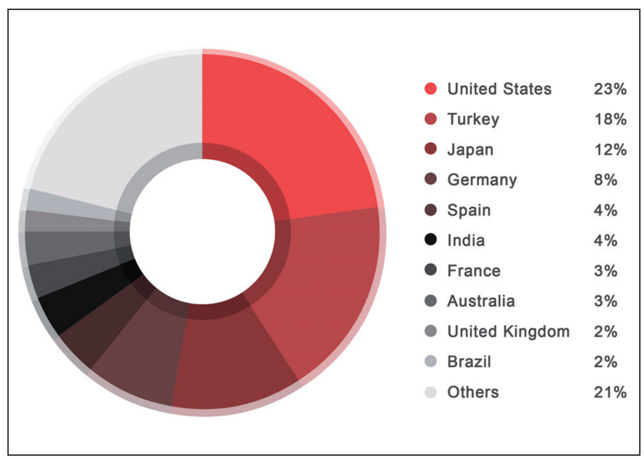

Figure 8: Distribution of clicks for Tweets with shortened URLs.

7.5 Traditional spam

The distribution of traditional spam attacks (shown in Figure 9) primarily focused on users in the US. We saw a large-scale health spam outbreak within the study period.

7.6 Twitter-specific scams

We discussed the impact of *Twitter* follower scams in Section 6.

7.7 Russian spam

The majority of users who clicked links embedded in Russian spam (shown in Figure 10) were from Russia (50%). However,

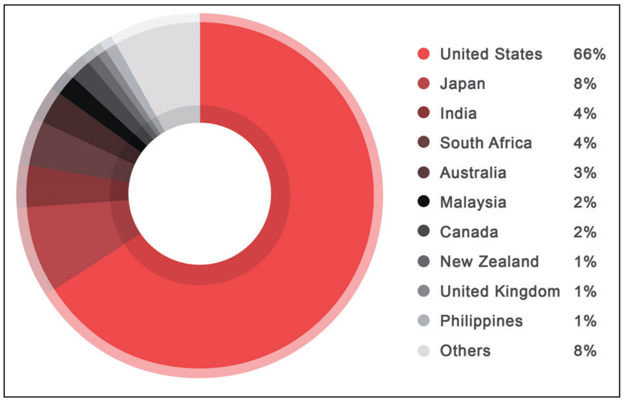

Figure 9: Distribution of clicks for traditional Twitter spam.

many users from non-Russian-speaking countries also clicked links in this kind of spam. We theorize that the contents advertised in this spam type (e.g. cracked software and games, free movies, cracks for mobile devices, exam and homework answers) appealed sufficiently to some users that they used automated translation tools to access inappropriate content.

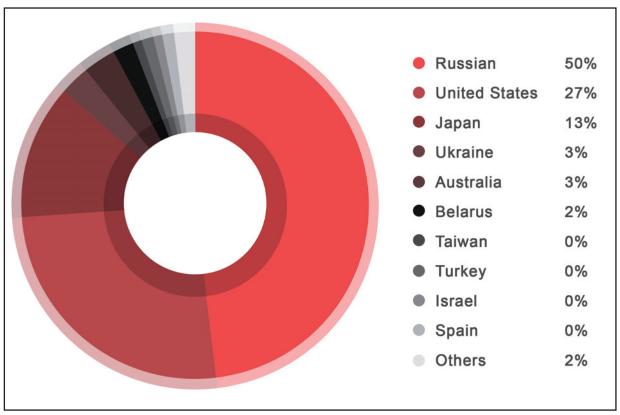

Figure 10: Distribution of clicks related to Russian spam.

8. IMPACT OF TWITTER PHISHING

In Section 2, we briefly described a *Twitter*-specific phishing scheme that has been going on for some years now [2]. We will discuss how such a scheme impacted *Twitter* and its users. This and similar schemes exploit the following features of *Twitter* in order to spread:

- They use URL shorteners.
- They have complex infection chains.
- The phishing Tweets were sent out via accounts that have been compromised.

In Figure 11, we considered the final page in the infection chain to be the 'phishing landing page.'

We approached this scheme from two angles – we determined how many posts on *Twitter* matched our phishing criteria and how many users attempted to load the phishing landing pages. We studied one particular scheme from March to May 2014.

Figure 11: Typical infection chain for a Twitter phishing scheme.

The largest outbreak we monitored occurred between 15–19 March 2014. On 18 March 2014, we identified 22,282 compromised users who sent out phishing Tweets with 13,814 distinct shortened URLs. On 19 March 2014, we identified 23,372 compromised users who sent out phishing Tweets with 5,148 distinct shortened URLs. The shortened URLs described here were confirmed to have infection chains that ended with phishing landing pages.

We tracked the number of users who landed on phishing websites within the study period and what countries they came from (see Figures 12 and 13). Throughout the study period, we noticed changes in cybercriminal tactics. From mid-March, we saw an ongoing attack develop into sporadic outbreaks in May. In March and April, the phishing landing pages had literal IP addresses as URLs, while the attacks in late May used more socially engineered host names using free web-hosting services.

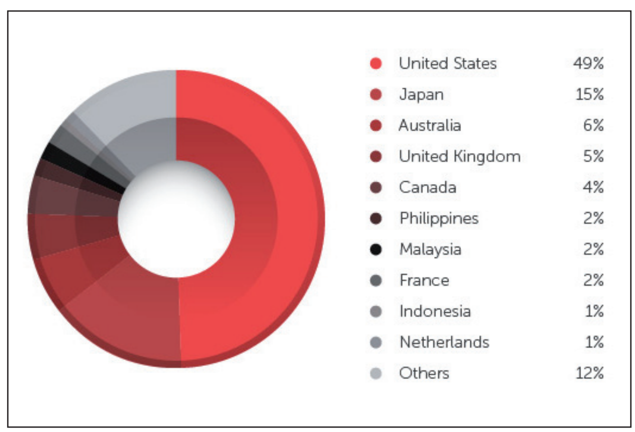

Figure 12: Distribution of users who attempted to access phishing landing pages.

CONCLUSION

This research paper presented a study of various types of abuse on *Twitter*. We analysed 500 million Tweets with embedded URLs and found that, during a period of high spam activity, 5.8% of them were spam or malicious in nature.

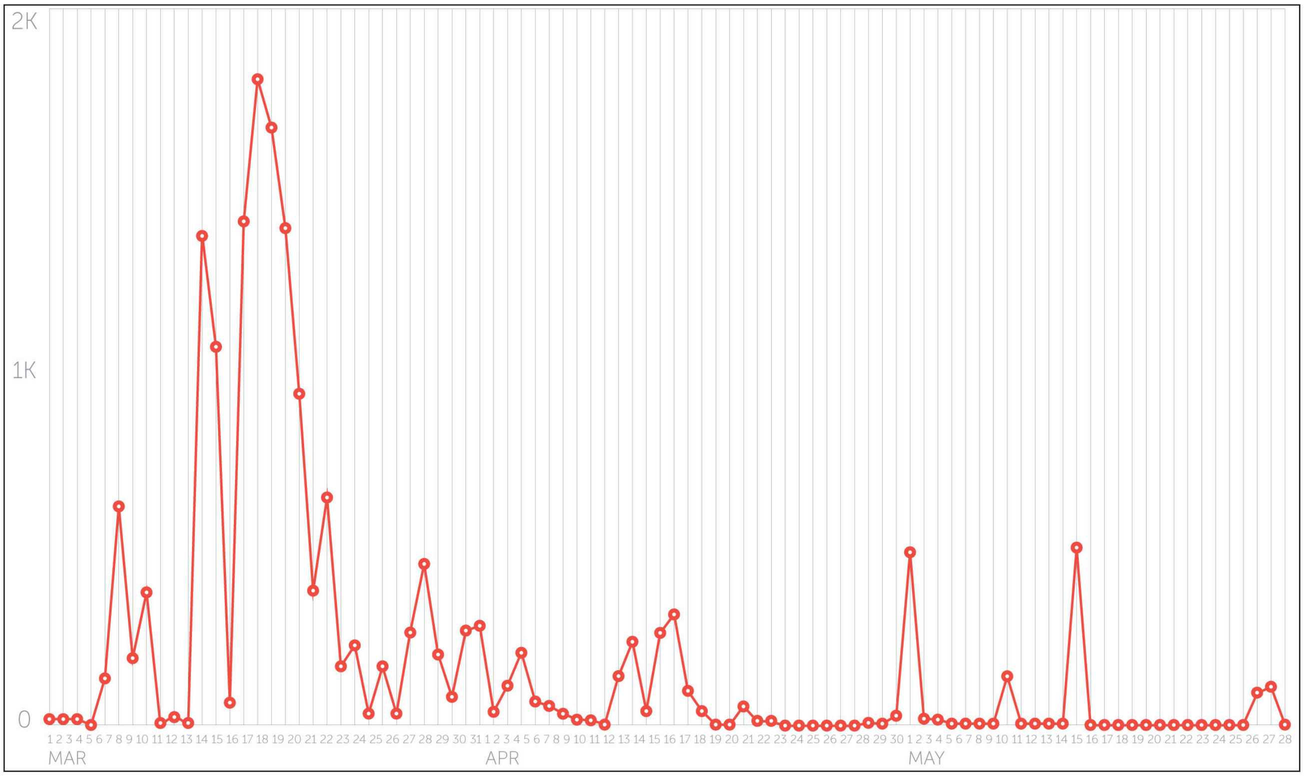

Figure 12: Number of users who attempted to access phishing landing pages.

We applied a hybrid technique, combining a blacklist augmented with algorithms suited for social networks, to the problem of identifying spam and malicious Tweets, which proved reasonably effective. The blacklist was augmented with a clique-discovery approach, which also very effectively identified large-scale spam outbreaks. We came to the conclusion that blacklists, when augmented in this way, are a useful tool in uncovering *Twitter* spam.

We examined the response rates for various types of *Twitter* spam and found that they varied widely, depending on the spam's content and other factors. We therefore conclude that quoting a single response rate for *Twitter* spam is inadequate; it is important to quote response rates for each type of spam instead.

We also examined the regional response rates for various *Twitter* outbreaks and found that they differed greatly across countries and regions.

ACKNOWLEDGEMENT

This work was supported by ARC Linkage Project LP120200266. We would like to thank Det Caraig and *Virus Bulletin* for assistance in preparing the manuscript.

REFERENCES

[1] Grier, C.; Thomas, K.; Paxson, V.; Zhang, M. @spam: The Underground on 140 Characters or Less. Proceedings of the 17th ACM Conference on Computer and Communications Security, pp.27–37. 2010. http://www.icir.org/vern/papers/ccs2010-*Twitter*-spam.pdf.

[2] Stone, B. Avoid 'Phishing' Scams. Twitter Blog. February 2010. https://blog.*Twitter*.com/2010/avoid-phishing-scams.

[3] Trend Micro Incorporated. Smart Protection Network – Data Mining Framework. 'Key Components'. 2014. http://cloudsecurity.trendmicro.com/us/technology-innovation/our-technology/smart-protection-network/#key-components.

[4] Wikipedia. Clique Problem. http://en.wikipedia.org/wiki/Clique_problem.

[5] Cheng, Y-C. Hadoop Success Stories in Trend Micro SPN. Hadoop in Taiwan Workshop 2012. http://www.gwms.com.tw/TREND_HadoopinTaiwan2012/1002download/04.pdf.

[6] Xiang, J.; Guo, C.; Aboulnaga, A. Scalable Maximum Clique Computation Using MapReduce. https://cs.uwaterloo.ca/~ashraf/pubs/icde13maxclique.pdf.

[7] Svendsen, M. S. Mining Maximal Cliques from Large Graphs Using MapReduce. Masters Thesis. http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=3631&context=etd.

[8] Pajares, P. Does the Twitter Follower Scam Actually Work? TrendLabs Security Intelligence Blog. 2014.

http://blog.trendmicro.com/trendlabs-security-intelligence/does-the-Twitter-follower-scam-actually-work/.

[9] Kanich, C.; Kreibich, C; Levchenko, K.; Enright, B.; Voelker, G. M.; Paxson, V.; Savage, S. Spamalytics: An Empirical Analysis of Spam Marketing Conversion. Proceedings of the 15th ACM Conference on Computer and Communications Security, pp. 3–14. 2008. http://www.icsi.berkeley.edu/pubs/networking/2008-ccs-spamalytics.pdf.

[10] Mindlin, A. Seems Somebody Is Clicking on That Spam. New York Times. http://www.nytimes.com/2006/07/03/technology/03drill.html?_r=2&.

[11] Catone, J. Spam ROI: Profit on 1 in 12.5m Response Rate. SitePoint. 2008. http://www.sitepoint.com/spam-roi-profit-on-1-in-125m-response-rate/.

[12] Zangerle, E.; Specht, G. "Sorry, I Was Hacked": A Classification of Compromised Twitter Accounts. 2014. http://www.evazangerle.at/wp-content/papercite-data/pdf/sac14.pdf.

[13] Egele, M.; Stringhini, G.; Kruegel, C.; Vigna, G. COMPA: Detecting Compromised Accounts on Social Networks. ISOC Network and Distributed System Security Symposium (NDSS). 2013. http://cs.ucsb.edu/~gianluca/papers/thjp-ndss13.pdf.

[14] Su, C-T.; Tsao, W-K.; Chu, W-R.; Liao, M-R. Mining Web Browsing Log by Using Relaxed Biclique Enumeration Algorithm in MapReduce. In Volume 3, pp.54–58, IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. http://www.computer.org/csdl/proceedings/wi-iat/2012/4880/03/4880c054-abs.html.