EDITORS: Rick Kuhn, US National Institute of Standards and Technology, kuhn@nist.gov
Tim Weil, Scram Systems, tweil.ieee@gmail.com

Spammers Are Becoming "Smarter" on Twitter

Chao Chen, Jun Zhang, Yang Xiang, and Wanlei Zhou,
Deakin University, Australia
Jonathan Oliver, *Trend Micro*

Twitter has become one of the most commonly used communication tools in daily life. With 500 million users, Twitter now generates more than 500 million tweets per day. However, its popularity has also attracted spamming. Spammers spread many intensive tweets, which can lure legitimate users to commercial or malicious sites containing malware downloads, phishing, drug sales, scams, and more.[1]

Spam is a problem throughout the Internet, and Twitter is not immune. In addition, Twitter spam is much more successful compared to email spam.[2] Various methods have been proposed by researchers to deal with Twitter spam, such as identifying spammers based on tweeting history[3] or social attributes,[4] detecting abnormal behavior,[5] and classifying tweet-embedded URLs.[6] Although researchers, as well as Twitter itself, have attempted to combat spam, the percentage of spam in the whole platform is still high. We hypothesize that this is because spammers are becoming more cunning on Twitter. While researchers are developing methods to detect spam, spammers continuously invent new strategies to bypass detection.

Here, we review the well-known methods that spammers have used to avoid or reduce their chances of being caught on Twitter. We show that spammers are now using more advanced strategies, namely *coordinated posting behavior, finite-state-machine-based spam template,* and *passive spam.*

Well-Known Spamming Strategies

At the most basic level, spammers use various Twitter functions such as @ and hashtags (#) to engage victims (see the "Twitter Features" sidebar for a breakdown of Twitter terms). Spammers can use @ to make spam tweets appear on the victim's feed without being a follower of the victim—for example, a spam tweet will appear on Barack Obama's timeline if it is written with @obama. By embedding popular hashtag keywords, one spam tweet can become part of a trending topic that can then be viewed by a victim who is interested in that topic. For example, a spam tweet with #007 will be disseminated to victims who are browsing the popular James Bond book and film series. Spammers also use other Twitter functions, such as "reply," "favorite," and "following" to spread spam.[7] Fortunately, researchers can also use these features (such as the number of followers or the number of hashtags) to detect Twitter spam.[3]

To bypass such detection systems, spammers apply evasion tactics, such as gaining more followers, posting more tweets, and so on.[8] They aren't exposed by the

 Published by the IEEE Computer Society

simple detection systems already described because their activity mimics that of legitimate users. To combat this, researchers propose robust social graph-based features, such as local clustering coefficients, betweenness centrality,[8] and distance/connectivity[4] to detect those tweets fabricated by spammers.

In addition to the aforementioned spamming strategies, our Twitter spam analysis reveals that spammers are now using more advanced methods.

Coordinated Posting Behavior

We collected a dataset of more than 570 million tweets with URLs from 25 September 2013 to 9 October 2013. Within this dataset, we identified around 33 million spam tweets using Trend Micro's Web Reputation Technology;[9] this accounts for 5.8 percent of the total tweets. We then used bipartite cliques to cluster the spam tweets into 17 groups, as shown in Table 1 (see the "Bipartite Cliques" sidebar for details about this method). Seventeen groups dominate more than 75 percent of the spam, whereas "others" account for less than 25 percent, indicating that, in general, spam is sent by groups.

We also found that six groups in Table 1 (groups A, B, C, E, I, and J, the bold, italicized letters in the table) had some common features:

- The URLs embedded in the tweets tend to use a .ru (server of Russian origin) domain.
- The content of the landing pages were written in Russian.
- The URLs tended to end with a Unix time stamp, such as http://xxxx.ru/xxxx-1380642617.html.

To study the spamming behavior of these six groups, we counted the tweets sent per hour by each group. Group A spread spam actively from 26 September (see Figure 1). When

Twitter Features

The following Twitter features were relevant to our work:

- Mention (@): If a tweet contains the @user tag, it is called a mention. The mentioned tweet can appear in a user's timeline even if the sender is neither followed by nor follows the user.
- *Hashtag* (#): A hashtag embedded in a tweet is normally a keyword to describe this tweet. If the hashtagged keyword is very popular, it will become a trending topic that can be seen by all Twitter users.
- *Timeline*: The tweets sent by those a user follows or tweets that use @ to mention the user appear in the user's timeline.

Table 1. Spam breakdown.

Group	Spam type	Spam tweets (%)
A	Edu and so on	27.28
B	Cracked software, games	8.11
C	Edu	6.26
D	Cracked	6.19
E	Cracked software	4.39
F	Printer, mobile	3.72
G	Twitter follower	2.54
H	Video, mobile, cracked software, games	2.23
I	Games, computer	2.04
J	Edu and so on	1.99
K	Shirt	1.91
L	Games, mobile printer	1.81
M	Computer, printer	1.77
N	Games, hardware spam	1.53
O	Computer game, mobile device	1.41
P	Credit-card, edu	1.08
Q	Cracked software, games	1.02
	Other spam	24.74

We named spam types according to the spam's content—for example, cracked software spam is about cracked software.

Bipartite Clique

To identify the bipartite cliques, we first extracted the domains of URLs embedded in tweets along with those tweets' senders. Then, we constructed a graph in which the Twitter users were nodes on one side of the graph while the domains in sent tweets were nodes on the other side. For each tweet from user *U* that contained a link with domain *D*, we connected this user *U* to domain *D* in the graph. Once the graph was fully connected, a bipartite clique was formed.

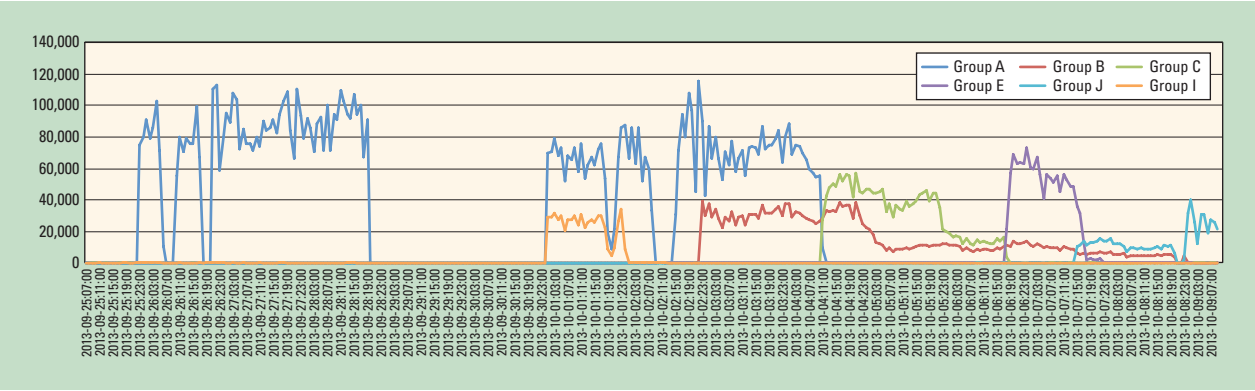

Figure 1. The number of spam tweets sent by the six groups highlighted in Table 1. (Note that data were lost for 29 and 30 September).

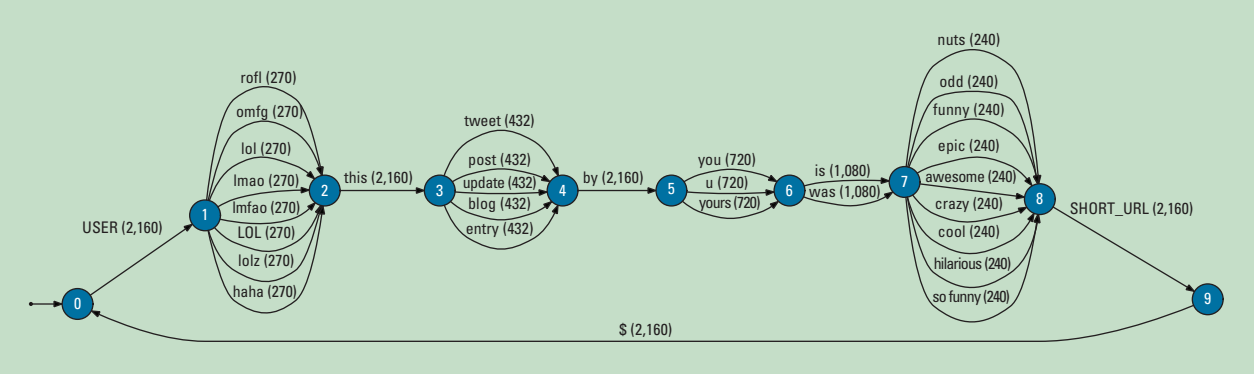

Figure 2. Finite-state machine-based spam template. Spammers can use this template to generate 2,160 different spam tweets.

group A stopped sending spam, group C started sending it on 4 October. Groups C and E, and E and J also displayed this type of spamming behavior. We regard this behavior as coordinated posting behavior, a phenomenon in which one group of spam tweets disappears while another is being sent. This kind of posting behavior is more difficult to detect because spammers change the groups of accounts to abuse Twitter.

Finite-State Machine-Based Spam Template

Some have found that most spam is generated using specific templates,[10] which is logical because it is very expensive for spammers to write each tweet manually. However, the template is often simple[10]—for example, "celebrity name" + "an eye-catching action" + URL. Therefore, researchers can extract the templates and match tweets to them to detect spam.

We found that spammers are now using more complex templates to generate spam. Surprisingly, spammers are using finite-state machines to generate what we have named finite-state machine-based spam templates (see Figure 2). One finite-state machine has a number of states, and each edge of it is denoted by one word. If we travel from the beginning to the end, we can have one full sentence, such as "lol, this tweet by you is funny + SHORT URL" in the finite-state machine. By using one finite-state machine-based spam template, spammers can generate many different tweets. Take the finite-state machine in Figure 2, for example; it has $8 \times 5 \times 3 \times 2 \times 9 = 2{,}160$ different routes from start to end. This means that spammers can use this template to generate 2,160 different spam tweets with little effort. For example, spammers can write a script that randomly chooses one option from each node to generate one spam tweet. Relying on simple string signatures to match spam tweets will allow most of these finite-state machine-based template spam tweets to escape detection.

Passive Spam

As previously described, traditional spam is distributed using Twitter functions such @ and #. However, we also found that much spam

does not use any tags. As a result, such spam cannot be identified by machine-learning-based spam detection that uses these features. Contrary to traditional spam, which tries to involve victims as much as possible, this spam is only viewed by victims when they search for specific key words. Consequently, we call this passive spam. None of these spam tweets have tags embedded (see Figure 3), and they are mostly promoting cracked games, software, or pirated movies.

We found that of the victims who clicked on this kind of spam,[9] 50 percent were in Russia. However, victims from many non-Russian-speaking countries also clicked on this kind of spam. Assuming these users did not speak Russian, we hypothesize that the content advertised in this spam was sufficiently enticing for victims to use translation software to access the inappropriate content. We also found that the suspended rate of this type of spam by Twitter is much lower than others, because spammers have much less interaction with users, allowing spammers to use this strategy successfully.

Although researchers and industry are devoted to developing detection and mitigation approaches to combat Twitter spam, spammers can thwart their efforts with ever-evolving techniques, such as the three complex spamming strategies we describe here. The war with spammers is becoming fiercer and is far from over; we should therefore continue to analyze spammers' behavior and propose robust spam-detection systems to make a safe Twitter environment for all users.

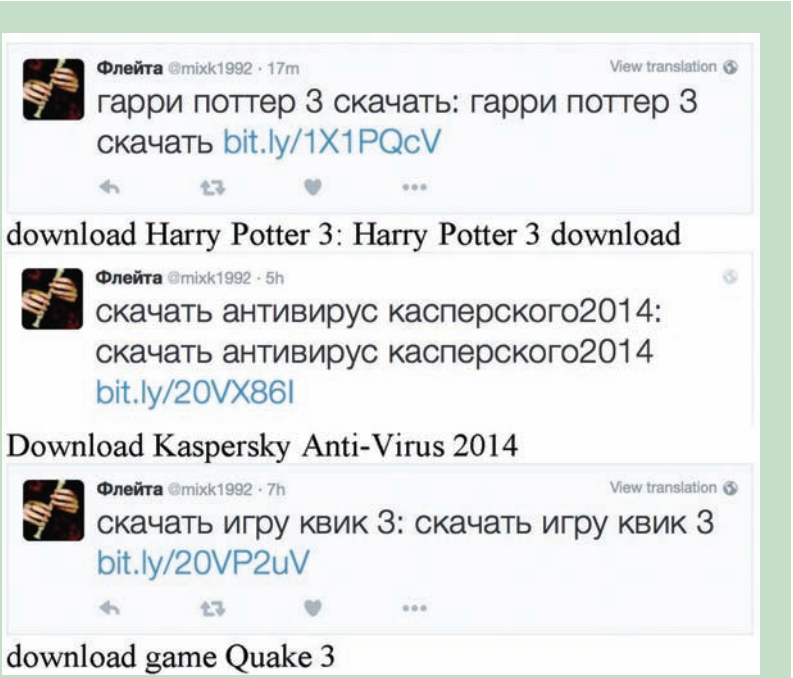

Figure 3. Examples of passive spam. Note that while many spam tweets originate from Russian domains, this strategy is also applied by spammers outside of Russia.

Acknowledgments

This work is supported by ARC Linkage Project LP120200266.

References

1. C. Chen et al., "6 Million Spam Tweets: A Large Ground Truth for Timely Twitter Spam Detection," *Proc. Int'l Conf. Communications*, 2015, pp. 7065–7070.
2. C. Grier et al., "@spam: The Underground on 140 Characters or Less," *Proc. 17th ACM Conf. Computer and Communications Security*, 2010, pp. 27–37.
3. F. Benevenuto et al., "Detecting Spammers on Twitter," *Proc. 7th Ann. Collaboration, Electronic Messaging, Anti-Abuse, and Spam Conf.*, 2010, http://www.decom.ufop.br/fabricio/download/ceas10.pdf.
4. J. Song, S. Lee, and J. Kim, "Spam Filtering in Twitter Using Sender-Receiver Relationship," *Proc. 14th Int'l Conf. Recent Advances in Intrusion Detection*, 2011, pp. 301–317.
5. M. Egele et al., "Compa: Detecting Compromised Accounts on Social Networks," *Proc. Ann. Network and Distributed System Security Symp.*, 2013, https://www.cs.ucsb.edu/~vigna/publications/2013_NDSS_compa.pdf.
6. S. Lee and J. Kim, "Warningbird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream," *IEEE Trans. Dependable and Secure Computing*, vol. 10, no. 3, 2013, pp. 183–195.
7. K. Thomas, "The Role of the Underground Economy in Social Network Spam and Abuse," PhD dissertation, Electrical Eng. and Computer Science Dept., Univ. of California, Berkeley, Dec. 2013.
8. C. Yang, R. Harkreader, and G. Gu, "Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers," *IEEE Trans. Information Forensics and Security*, vol. 8, no. 8, 2013, pp. 1280–1293.
9. J. Oliver et al., *An In-Depth Analysis of Abuse on Twitter*, tech. report, Trend Micro, Sept. 2014; www.trendmicro.com/cloud-content/us/pdfs/security-intelligence/white-papers/wp-an-in-depth-analysis-of-abuse-on-twitter.pdf.
10. H. Gao et al., "Spam Ain't as Diverse as It Seems: Throttling OSN Spam with Templates Underneath," *Proc. 30th Ann. Computer Security Applications Conf.*, 2014, pp. 76–85.

Chao Chen *is working toward a PhD in computer science at the School of Information Technology, Deakin University, Australia. His research interests include network security and social network security. Chen received a BS in information technology (with first class honors) from Deakin University. Contact him at chao.chen@deakin.edu.au.*

Jun Zhang *is with the School of Information Technology, Deakin University, Australia. His research interests include network and system security, pattern recognition, and multimedia processing. Zhang received a PhD from the University of Wollongong, Australia. Contact him at jun.zhang@deakin.edu.au.*

Yang Xiang *is the director of the Centre for Cyber Security Research and a professor at the School of Information Technology, Deakin University, Australia. His research interests include network and system security, distributed systems, and networking. Xiang is the chief investigator of several projects in network and system security, funded by the Australian Research Council (ARC). He received a PhD in computer science from Deakin University. Contact him at yang@deakin.edu.au.*

Wanlei Zhou *is the chair professor of IT at the School of Information Technology, Deakin University, Australia. His research interests include network security, distributed and parallel systems, bioinformatics, mobile computing, and e-learning. Zhou received a PhD from the Australian National University and a DSc from Deakin University. Contact him at wanlei.zhou@deakin.edu.au.*

Jonathan Oliver *is a senior architect at Trend Micro, where he focuses on anti-spam and Web reputation technologies. He performed postdoctoral research in Australia and the UK, and acted as a data-mining consultant in Silicon Valley. Oliver led the anti-spam R&D at Mailfrontier, an anti-spam start-up, from 2002 to 2006. He received a PhD in computer science from Monash University, Australia. Contact him at jon_oliver@trendmicro.com.*