Observing Common Spam in Tweets and Email

Cristian Lumezanu
NEC Laboratories America
Princeton, NJ
lume@nec-labs.com

Nick Feamster
University of Maryland
College Park, MD
feamster@cs.umd.edu

ABSTRACT

Spam is pervasive across many types of electronic communication, including email, instant messaging, and social networks. To reach more users and increase financial gain, many spammers now use multiple content-sharing platforms—including online social networks—to disseminate spam. In this paper, we perform a joint analysis of spam in email and social networks. We use spam data from Yahoo's web-based email service and from Twitter to characterize the publishing behavior and effectiveness of spam advertised across both platforms. We show that email spammers that also advertise on Twitter tend to send more email spam than those advertising exclusively through email. Further, we use DNS lookup information to show that sending spam on both email and Twitter correlates with a significant increase in coverage: spam domains appearing on both platforms are looked up by an order of magnitude more networks than domains using just one of the two platforms.

Categories and Subject Descriptors

C.2.3 [**Computer-communication networks**]: Network Operations; K.4.2 [**Computers and society**]: Social issues; H.0 [**Information systems**]: General

General Terms

Measurement, Security

Keywords

Twitter, email, spam, DNS, multiple platform spam

1. INTRODUCTION

Spam is an unwanted yet continual scourge in electronic communication; it infiltrates basic conversation tools, such as email and instant messaging, as well social-based content sharing platforms, such as OSNs and forums. A recent study shows that 89% of the 107 trillion emails sent in 2010 were spam [15]. Further, emerging social-based applications such as social networks, blogs, or video sharing websites, are also subject to abuse: 8% of the URLs shared on Twitter lead to malicious websites that host malware or scams [6].

To reach more users and maximize their financial gains, spammers are increasingly using multiple content-sharing platforms to disseminate the same malicious content. For example, several malware and phishing campaigns are carried over email, to target as many users as possible, but also over social networking platforms, where the trust-based social graph enables spammers to target those users who are more likely to click on their links [8,9]. In our analysis (§ 4), we discover over 700 scam, malware, and phishing domains that are advertised using both email and tweets over the course of a month.

In this paper, we take an initial step towards understanding the properties of spam across multiple platforms. We present a joint measurement study of spam across two popular content sharing platforms: email and social networking. We consider spam to be any message containing an URL that leads to a website that hosts malicious content such as malware, phishing, or scams. Using two message data sets, of emails to Yahoo accounts and tweets from Twitter, gathered for one month in March 2011, we provide a parallel view of the properties of spam and behavior of spammers. An important characteristic of our study is that it focuses on *both emails and tweets sent during the same time period.*

We characterize email and tweet spam from two different perspectives. First, we investigate the presence of spam on the two dissemination platforms (Section 4). We discover that 55% of all spam emails advertise scam domains that also appear on Twitter. Email spammers that advertise at least one domain on Twitter are likely to send more email spam overall during our measurement than those that send only email-exclusive spam.

Second, we seek to understand the effect of advertising spam on multiple platforms (Section 5). To the best of our knowledge, our study offers the first large-scale analysis on the behavior and effectiveness of spam sent during the same time period on both email and Twitter. Using DNS lookup data to estimate the number of clicks for each domain, we find that domains common to both emails and tweets receive an order of magnitude more clicks than domains exclusive to one platform.

Because the email data set is unfiltered while the tweets have already passed through Twitter's spam filters, our study has a few limitations. We cannot offer a complete comparison between email and tweet spam or establish a causal

IMC'12, November 14–16, 2012, Boston, Massachusetts, USA.
Copyright 2012 ACM 978-1-4503-1705-4/12/11 ...$15.00.

relationship between the presence of spam in tweets and in emails and its effectiveness. Notwithstanding, our results show that there is a positive correlation between publishing spam on multiple platforms and both the volume of email spam send by these publishers and the effectiveness of the spam being sent. This leads to a most important finding of our study: *incorporating information about Twitter spam publishing behavior into email spam filters (or vice versa) could increase their effectiveness in quickly identifying virulent spammers.*

Although we do not pursue the positive implications of our findings in this paper, we believe that the overlap between spam that appears in both email messages and tweets ultimately presents new opportunities for improving both the accuracy and speed of spam detection. Spam filters could leverage information about publishing behavior of spam across platforms to build better defenses. For example, discovering that spam URLs are consistently published on Twitter before email could help develop more accurate email blacklists. Further, understanding how and when spammers send common spam could reveal their strategies and how to counter them. Finally, understanding the presence and behavior of common spam would help existing spam fighting solutions developed for individual platforms [1, 2, 5–7, 10, 12, 16, 19] interact better with each other.

2. RELATED WORK

Most analyses of spam have focused exclusively on email [1,11,12,19]. They characterize spam from various aspects such as the behavior of spammers [11, 12], the email contents [14], or the properties of spam hosting infrastructures [1]. More recently, several studies characterize spam in forums [10, 13], or in social networking applications such as Facebook and Twitter [2, 5–7]. Unlike previous work, we focus on spam that is sent on *multiple platforms (e.g., both email and Twitter) at the same time*. Such a joint analysis provides a new perspective on spam and allows us to better understand how spammers work and coordinate to increase their impact.

To fight spam sent across web services such as social networking or video sharing sites, Thomas *et al.* proposed Monarch, a spam URL filtering service [16]. Monarch compared email and tweet spam and found few features in common across the two platforms. This means that one would need to learn specific sets of rules to detect spam on each platform. We share a similar philosophy to Monarch, that generalizing spam fighting across multiple platforms can lead to faster and more efficient detection. However, unlike Thomas *et al.*, we focus on the spam that *is common across platforms* and characterize its prevalence, publishing behavior, and effectiveness.

3. DATA AND METHODS

In this section, we present the data sets used in our analysis and discuss how we identify spam messages. We use Twitter and Yahoo! Mail as representative applications for online social networks (OSNs) and email, respectively.

3.1 Data sets

To compare spam in tweets and email, we start with two sets of public messages, *Twitter* and *Yahoo*, both collected in March 2011. The *Twitter* data set contains public tweets gathered using the public Twitter streaming API. For each tweet we retrieve the publishing time, details on its author (*e.g.*, id, screen name, numbers of followers, friends, and statuses posted), as well as the text of the tweet. The *Yahoo* data set represents a snapshot of all incoming emails to Yahoo! Mail accounts as received by Yahoo's mail servers and it was privately shared by Yahoo!. The information available is the connecting IP, the time of the message, and the URLs contained in the message. Both data sets capture around 1%, extracted uniformly at random, of all messages sent on each platform in March 2011.

3.2 Spam identification

Because URLs are the primary method that spammers use to attract users to websites that host malicious content, we restrict our analysis on the tweets and emails that contain URLs. To identify spam messages, we take the following three steps. (1) We parse the text of each tweet and extract all URLs. (2) We find the final landing page for all links from tweets and emails that are hidden with URL shortening services or behind chains or redirections. (3) We use two sources to determine whether a URL leads to malicious content by comparing the URL of the final landing page with public URL blacklists and spamtrap emails. We describe below this final step in more detail.

URL blacklists contain domains and websites that are known to host malicious content such as scams and malware or participate in phishing campaigns. We check the domains in our data sets against several public URL blacklists in the first week of May 2011. URIBL and SURBL are DNS-based blacklists with domain names found in the body of spam emails, but generally not in legitimate emails. PhishTank lists phishing URLs voted by users. To detect malware, we use the list published by `malwaredomains.com`. Finally, for both malware and phishing domains, we use the Google Safebrowsing API to check against Google's constantly updated blacklists.

Blacklists are not always effective in detecting recently advertised spam domains or domains that are not frequently reused. Previous studies show that information about many spam domains and spammers fails to show up in blacklists even more than a month after the domain was first advertised [6, 12]. To improve the completeness of our study and detect even recently published spam domains, we use data from the spamtrap set up by Ramachandran and Feamster [11]. Because the spamtrap is associated with a DNS Mail Exchange (MX) record with no legitimate email addresses, all email that it receives is spam. For our measurement, we collect all URLs found in over 11 million emails received by the spamtrap between January and March 2011.

Although all emails received at the spamtrap are from spammers, the spam emails might still contain legitimate URLs to subvert anti-spam filters. We use the following simple heuristic to select the spamtrap URLs (and domains) that are more likely to be malicious. First, we whitelist all URLs with domains present in the Alexa top 10,000 most popular domains, under the assumption that popular domains rarely host spam. Second, we consider only domains that appear in more than 1,000 spamtrap emails. This is based on the assumption that, for spam to be effective, it must be distributed at scale [20].

Data set	Messages with URLs			
	All	Spam		
		blacklist	spamtrap	total
Yahoo	290,355,683	39,440,693	13,134,298	49,142,499
- Yahoo only	135,740,285 (47%)	29,755,924 (75%)	8,012,716 (61%)	34,159,814 (69%)
- Yahoo, common	186,552,390 (64%)	18,712,383 (47%)	8,881,790 (68%)	27,009,682 (55%)
Twitter	5,569,940	197,932	18,404	198,887
- Twitter only	496,303 (1%)	29 (<1%)	0 (0%)	29 (<1%)
- Twitter, common	5,159,890 (99%)	197,903 (>99%)	18,404 (100%)	198,858 (>99%)

Data set	Domains			
	All	Spam		
		blacklist	spamtrap	total
Yahoo	14,860,901	81,567	3,681	82,233
- Yahoo only	14,699,376 (>99%)	80,897 (99%)	3,601 (89%)	81,493 (99%)
- Yahoo, common	161,535 (< 0%)	670 (<1%)	80(2%)	740 (1%)
Twitter	283,936	676	80	746
- Twitter only	127,315 (45%)	6 (<1%)	0 (0%)	6 (<1%)
- Twitter, common	156,621 (55%)	670 (>99%)	80 (100%)	740 (>99%)

Table 1: Data sets used in our study. We collect the main data sets, *Yahoo* and *Twitter* in March 2011. The secondary data sets contain the messages with domains that appear only on Yahoo, only on Twitter, and on both. Percentages are computed of the values for the main data sets. Not all percentages add to 100% because there are messages with more than one URL where some URLs are unique to one platform and others are common to both. We also separate the messages and domains according to whether they are spam or not and to the method used to identify them as spam (blacklists or spamtrap).

3.3 Spam characterization

We use the methods described above to identify spam messages in the *Twitter* and *Yahoo* data sets. Table 1 shows the statistics with respect to spam domains and messages (tweets or emails) that contain spam URLs (focus on the data sets labeled "Yahoo" and "Twitter"). There are around two orders of magnitude more spam emails than tweets. This could be explained by the relative difference between the number of messages in the two data sets but it could also be due to the Twitter data being collected after filtering, as we explain below. The gap becomes smaller when we consider the amount of spam relative to the total number of messages on each platform: 17% of all emails are spam while, even after filtering, 4% of all tweets are spam. These numbers suggest that although Yahoo! Mail has many more users and daily messages than Twitter [4, 18], Twitter is becoming a significant infrastructure for disseminating spam.

3.4 Limitations

Data sampling. Due to the large volume of tweets and emails, we analyze only 1% of the emails received by Yahoo servers and of the tweets sent in March 2011. Sampling can misrepresent the properties of the original message population, but we are confident that it has not distorted general trends: for example, the statistics about the prevalence of spam that we report are comparable to what previous studies have shown [6, 17, 20].

Spam filtering. The *Yahoo* data set contains information about emails captured *before* any spam filtering was performed, while the tweets in our data set have already passed through Twitter's spam filters. This can affect our results as follows:

1. It can underestimate the amount of Twitter spam, which in turn underestimates the common spam sent on both Twitter and email. However, even after filtering, we find a significant amount of tweets that are spam.
2. Twitter's filtering may skew the ratio between the amount of "common with Yahoo" and "exclusive to Twitter" spam (if exclusive domains are filtered more or less than common domains). We believe that, because Twitter uses email blacklists to fight spam [6], it is more likely that common domains are filtered more than exclusive domains, making our results in Section 4 an underestimation of reality.
3. Finally, while distorting the absolute view on common spam, the filtered tweets in our data sets can offer a more realistic perspective on the amount of common spam that *an email spam filter would see* if it had access to the publicly available Twitter stream. Our results in Section 4 suggest that email filters could be more effective if they incorporated Twitter spam publishing behavior in their analysis.

Spam identification. We use both blacklists and spamtrap emails to ensure that our knowledge about which domains host malicious content is as complete as possible. However, even then, we may still miss many spam domains. For example, neither blacklists nor spamtrap emails help us identify several Twitter-specific spam campaigns (*e.g.*, phishing for followers, buying retweets [6]). Because the focus of our paper is on common spam in email and Twitter, missing the Twitter-specific spam does not affect our results.

4. PRESENCE OF COMMON SPAM

We now study the presence of spam and the behavior of spammers across Twitter and email from two perspectives: How is common spam published across platforms?, and How do we identify it? Unsurprisingly, we find that the networks

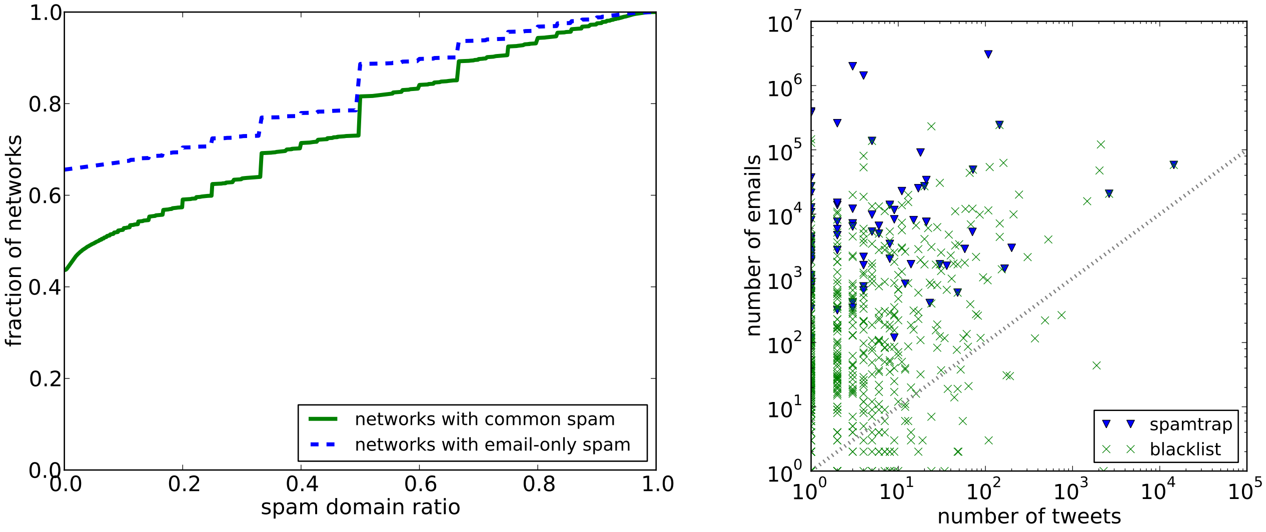

Figure 1: (left) Cumulative distribution for the spam domain ratio for /24 networks that send spam; the spam domain ratio for a network is the fraction of the number of spam domains to the total number of domains sent from them network; (right) Number of emails/tweets for each common spam domain; each point is associated to a common spam domain and has a different symbol according to how the domain was detected.

that advertise spam on both email and Twitter tend to send more spam overall than the networks that advertise spam exclusively through email. We also study the sources that allow us to identify spam and show that URL blacklists help identify the most spam, although spamtraps are also effective to discover high-volume spam domains.

We divide each data set into two secondary data sets, according to whether a message contains URLs that are published on both platforms or are exclusive to one platform. Table 1 presents details on the secondary data sets. There are 740 spam domains common to both Yahoo and Twitter which appear in 27M emails and 198K tweets. More than half (55%) of the spam emails contain domains that appear on Twitter. Because the tweets in our data set have already passed through Twitter's spam filters (as explained in Section 3.4), we cannot make a definitive assessment on the amount of common spam. However, because Twitter uses email blacklists to filter spam [6], if tweets were not filtered, we would likely have even more spam domains in common with email. Thus, the numbers in Table 1 could be interpreted as lower bounds for the amount of common spam.

4.1 Publishing behavior

We focus on the common spam domains that we do identify. We compare how much each domain is advertised across the two platforms. Only 10% of the common domains appear in more tweets than emails and 60% of the common domains appear in at least ten times as many emails as tweets. We also find that 99% of all domains that appear in tweets also appear in emails. These results could be an artifact of the pre-filtering of tweets (*e.g.*, if Twitter had already filtered the common spam that is more prevalent in tweets) or of using email-based blacklists to detect Twitter spam. However, if we take into account the total number of messages sent on the two platforms, the result could also mean that email is still a more pervasive platform for sending spam and that Twitter may be used simply as a backup for email. To better understand the implications of our observation, we are currently implementing Twitter-specific spam detection techniques [3] that are trained independently from email spam.

Next, we compare the spamming behavior of users that send spam on both platforms to that of users that send spam on only one platform. In doing so, we hope to understand whether common spam domains are associated with heavier spammers. We restrict our analysis to Yahoo users because we find little Twitter-only spam for a meaningful comparison. We extract the /24 networks from which every email was sent and group them into two categories: those from which at least one common spam domain was advertised and those from which no common spam, but at least one email-only spam domain, was sent. For each network, we compute the spam domain ratio as the fraction of spam domains to the total number of domains sent from the network.

The left graph in Figure 1 shows the distributions of the spam domain ratio for networks that originate common spam and networks that are never source for common spam. The networks that advertise at least one common spam domain have a higher ratio than those that do not. Although this observation does not necessarily imply a causal relationship between sending spam on multiple platforms and the volume of email spam, it does indicate a correlation that could be exploited by spam filters to improve their efficiency: networks that send spam domains that also appear on Twitter are likely to send more spam overall than those networks advertising domains that are exclusive to email.

4.2 Identification

Given that detecting common spam domains can help identify networks that are likely to send more spam, we ask what is the better source for detecting these domains.

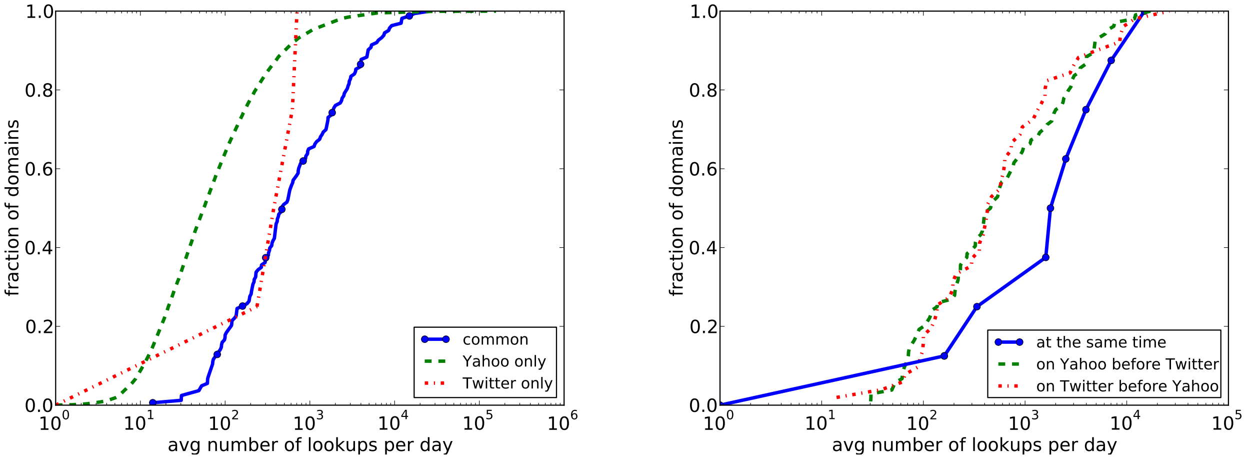

Figure 2: Cumulative distribution of the average number of unique networks querying for spam domains per day: (left) all spam domains, and (right) common spam domains.

We separate the spam messages according to the information source (blacklists or spamtrap) used to classify them as spam. Table 1 shows statistics about the results. Only 7% of all spam emails and 9% of all spam tweets could be identified using *both* sources, which is evidence for the lack of overlap between our spamtrap data and URL blacklists. Although spamtraps constitute an important source of information for building blacklists, oftentimes blacklists are not updated quickly enough to keep up with the "fresh" spam [12].

Although spamtraps appear to detect fewer spam domains than blacklists (4% of all email spam domains are identified using the spamtrap data), these domains are significant on email: they contribute to 27% of all email spam and to 33% of all common spam sent through emails. For Twitter, spamtrap data is less decisive, identifying only 9% of all tweets. For further evidence, the right graph of Figure 1 shows the number of tweets vs. the number of emails each common domain appears in. Each point is associated with a domain and has a different symbol according to whether the domain was identified as spam using blacklists or spamtrap data. In conclusion, *blacklists can identify most common spam, although spamtrap data is also effective in detecting the common domains that appear in many emails.*

5. EFFECTIVENESS OF COMMON SPAM

We presented evidence that there is a significant amount of common spam across tweets and emails: 55% of all spam emails and 99% of all spam tweets advertise content also published on the other platform. We seek now to better understand the role of each platform in the spammers' strategies. Is Twitter a backup for cases when email spam is not effective? Or do the platforms combine to capture a more diverse set of users, unattainable from a single medium?

To understand the effectiveness of spam on each platform, we use DNS lookup information collected by Verisign. Verisign manages the TLD nameservers responsible for the `.com` and `.net` domains. We obtained a data set with information about when each domain (ending in `.com` or `.net`) in our data set was looked up. We consider only those lookups performed during the time when our tweets and emails were sent (March 2011). The data set contains information about the /24 networks of recursive resolvers that looked up 66% of the common spam domains, 57% of the Yahoo exclusive domains, and 66% of the Twitter exclusive domains. To estimate the number of clicks on each domain, and implicitly its popularity, we use the number of unique /24 networks from which a DNS query is performed.

Figure 2(left) presents the distribution of the number of unique networks that lookup each spam domain in March 2011. There are around ten times more networks that lookup common spam domains than spam domains that are exclusive to email. We cannot make a definitive assessment for the Twitter exclusive domains due to their low number. We further separate the common domains according to when they were first advertised: on both platforms on the same day, on Twitter before email, and on email before Twitter. We plot the distributions for each category in Figure 2(right). The domains that are published for the first time on both platforms on the same day tend to receive more clicks than those that are published on different days.

We find that there is a positive correlation between publishing spam on multiple platforms and its effectiveness, measured as the number of unique networks that lookup the spam domains. Our data is insufficient to determine whether the number of dissemination platforms or some other unobserved property of the spammer (*e.g.*, some spammers may be sophisticated enough that they send high yield spam *and* use multiple dissemination platforms) is responsible for the increased effectiveness of spam. Notwithstanding, the association that we observe strengthens or finding from Section 4.1, that sharing spam information across platforms (*e.g.*, whether a spam domain appears on both Twitter and email) could help email spam filters detect the heavy and virulent spammers quicker.

Equating the number of clicks that a domain receives with the number of unique /24 networks from which DNS queries are performed for it can introduce bias in our results. Most importantly, due to caching of DNS information along the DNS hierarchy, not all clicks on an URL lead to queries to the TLD nameservers. However, we believe that the number of unique networks still reflects the *relative* popularity between domains. If anything, it may under-represent the more popular domains, for which it is more likely that a query is cached at a lower level resolver.

6. CONCLUSIONS

We presented a measurement study on the properties of common spam across multiple content sharing platforms. We focused on two popular web applications, Twitter and Yahoo! Mail. We make two main observations: (1) spam sent on both Twitter and email at the same time has a better exposure and a higher lookup rate than spam sent exclusively with email; and (2) spammers that advertise on both on email and Twitter send more email spam overall than spammers that advertise only on email.

The limitations of the data sets do not allow us to draw definitive conclusions about how spammers use Twitter and email. For example, we cannot determine whether there is causation or simply correlation between sending spam concurrently using multiple mechanisms and its virulence and volume. Notwithstanding, our results suggest that general solutions for detecting message abuse that incorporate information and features from multiple platforms at the same time may ultimately improve the accuracy and responsiveness of existing filters.

7. ACKNOWLEDGEMENTS

We thank our shepherd Christian Kreibich and the IMC reviewers for their valuable feedback. This project was supported by NSF grants CNS-0643974 and CNS-1111723, and by NSF grant 0937060 to the Computing Research Association for the CIFellows Project.

8. REFERENCES

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

[2] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on Twitter. In *CEAS*, 2010.

[3] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *CEAS*, 2010.

[4] Full metal email: Confessions of an 'anti-spam zealot'. `http://goo.gl/WL8IG`.

[5] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaign. In *ACM IMC*, 2010.

[6] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The underground on 140 characters or less. In *ACM CCS*, 2010.

[7] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots + machine learning. In *SIGIR*, 2010.

[8] Newest messaging malware targets Facebook and Twitter. `http://goo.gl/TrGkq`.

[9] Unstoppable new phishing attacks blanket Facebook, Twitter, Hotmail. `http://goo.gl/X0tdr`.

[10] Y. Niu, Y. min Wang, H. Chen, M. Ma, and F. Hsu. A quantitative study of forum spamming using contextbased analysis. In *NDSS*, 2007.

[11] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. In *ACM Sigcomm*, 2006.

[12] A. Ramachandran, N. Feamster, and S. Vempala. Filtering spam with behavioral blacklisting. In *ACM CCS*, 2007.

[13] Y. Shin, M. Gupta, and S. Myers. Prevalence and mitigation of forum spamming. In *IEEE Infocom*, 2011.

[14] Spam forensics: Reverse-engineering spammer tactics. `http://goo.gl/Y9wmk`.

[15] Symantec messagelabs intelligence: 2010 annual security report. `http://www.messagelabs.com/mlireport/MessageLabsIntelligence_2010_Annual_Report_FINAL.pdf`.

[16] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and evaluation of a real-time URL spam filtering service. In *IEEE Security & Privacy*, 2011.

[17] State of twitter spam. `http://goo.gl/M4X5M`.

[18] Twitter finally reveals all its secret stats. `http://goo.gl/07kzP`.

[19] Y. Xie, F. Yu, K. Achan, E. Gillum, M. Goldszmidt, and T. Wobber. How dynamic are ip addresses? In *ACM Sigcomm*, 2007.

[20] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming botnets: signatures and characteristics. In *ACM Sigcomm*, 2008.