Summary Review Documentation for

“Observing Common Spam in Tweets and Email”

Authors: C. Lumezanu, N. Feamster

Reviewer #1

Summary: This paper compares spam on Twitter and Yahoo Mail. The authors use two datasets collected in March 2011, one of tweets collected from the streaming API (with metadata including author, number of followers, etc, and after spam-filtering) as well as a sample of SMTP messages arriving at Yahoo’s SMTP servers, including timestamp, sending IP address, and URLs in the message (pre-spam-filtering). Each dataset consists of 1% of the full message volume on either platform. The authors extract URLs for the messages in either dataset and crawl them, checking the landing page URL against blacklists including SURBL, URIBL, PhishTank, malwaredomains.com, and Google Safebrowsing. They also include a homebrew blacklist produced by running a spam trap that collected 11M messages between January and March 2011. They whitelist from the top Alexa domains.

They discuss limitations, the largest of which seems to be the fact that the tweets get collected post-spam-filtering whereas no filtering has been applied to email, resulting in dataset asymmetry. For domains appearing in both tweets and email, they find that the volume on email far surpasses that in tweets. Virtually all tweeted domains also appear in email. They find that cross-spamming of a domain is an indicator of stronger email spam activity than seen for email-only spam.

They find blacklists work well for identifying spam in both media, whereas the domains identified via spambots reflect large parts (27% for email, 33% for all spam) of the total spam volume. Using DNS lookups recorded at Verisign, the compare the lookup volumes of spam domains for the platforms. Common domains are looked up roughly 10x more often than email-only domains.

Strengths: The implications of spammers leveraging multiple platforms are interesting, and it’s important to understand emerging trends in this area.

Weaknesses: The filtering asymmetry in the collected datasets makes meaningful comparison difficult. The argument from 4 that Twitter filters using blacklists, doesn’t really help because you don’t know volume, distribution, nor blacklist timeliness of tweets that may get filtered this way.

Comments to authors: Given that your Twitter dataset contains post-filtered messages whereas Yahoo’s is pre-filtering, I feel the comparisons are inherently apples-to-oranges. The most meaningful part seems to be the spambot-only subset of the datasets, as you could reasonably argue that blacklists helped neither tweets nor email, but as no such domains appear exclusively on Twitter this doesn’t help much. It might be fun to try to reverse-engineer the blacklist composition Twitter employs, from the domains you see blocked by it (or not) over time.

Additional comments:

• In 3.1 it would help to get baseline numbers for your datasets.

• 3.2 suggests that your URL crawling may have side effects for Yahoo Mail’s users, as your dataset contains non-spam messages.

• 3.3: only 17% of all emails in your dataset are spam? That seems really quite low.

• Table 1 would benefit from consistent use of percentages so one can easily understand falloff as one proceeds through the table.

• In 5, what is the role of crawlers in your DNS data? Do you look for them? Are they significant?

• Finally, you should acknowledge that Monarch *did* compare email and Twitter spam, if only on half a page. Nevertheless, they found little overlap of classification features between the two classes, so you should comment on that.

Reviewer #2

Summary: This paper examines the relationship between the URLs that appear in spam that originates from email and spam that appears as tweets on Twitter. The authors collect data from both Twitter and Yahoo for a period of one month, and look for URLs that appear in messages that also appear in blacklists and spam traps. The authors find that spam that appears in both email and Twitter appears to have higher volume and also appears to originate from more sources.

Strengths: The paper looks at a hot topic (spam in Twitter) and takes a new approach by comparing it to the more-well-understood problem of spam in email. The paper uses a nice data set (Yahoo email, as well as DNS requests).

Comments to authors: Overall, I enjoyed reading this paper and think it is an interesting direction to pursue for anti-spam research. I especially liked the data sets that you used. However, I have a few concerns about this paper, outlined below.

• My primary concern is over your classification of spam in Twitter. You are essentially using email-based classification methods to determine whether things are spam. As a result, it is somewhat unsurprising at you find much more spam in email that you find spam in tweets. Much of the Twitter spam
literature that I’ve read uses alternate methodologies for detecting Twitter spam (most notably, looking for banned accounts two weeks later). I’m wondering how much different your results would be if you had used a Twitter-specific methodology for spam collection than an email-specific one.

To wit, it would seem that many of your statements are a bit too strong given this limitation. For example, “email is still a more pervasive platform for sending spam” – this may be different if there are many Twitter spams that you’re missing.

• I was somewhat confused by your use of the spam trap in 3.2. First, your methodology for constructing the spam trap is extremely under-specified. How did you create it? How did you advertise the addresses? How much of the email was received within the same time period as your Yahoo/Twitter data? Second, I’m a bit unclear about the motivation for the spam trap. Wouldn’t the blacklists that you use be much more comprehensive than your spam trap? Presumably, Google’s list would be much more highly curated and accurate.

• You don’t describe how the Yahoo messages are extracted from the full Yahoo stream. Are they a 1% random sample (ala the Twitter spritzer)? If not, could there be any bias there?

• Overall, your conclusions are not particularly surprising – domains that appear in both Twitter and email are likely to be sending more spam, from more places, and getting more clicks. I’m wondering if you couldn’t look at more interesting issues, for example, what kinds of domains appear in both vs. just one? Are the same exact URLs advertised? Since twitter users must have accounts, do you see correlations between accounts sending multiple domains and those domains in email spam?

Reviewer #3

Summary: The paper tries to study spam by looking at the common spam campaigns across emails and social network spam. The authors look at twitter and yahoo mail, and analyze common campaigns found across the two datasets. Overall, the paper was unable to draw many conclusions due to the nature of the data, and most conclusions came with caveats and qualifications. Ultimately, I did not learn anything significant through the paper.

Strengths: The problem of analyzing spam across email and social networks has to my knowledge never been done before. The datasets are significant, particularly the yahoo email dataset, which is not easy to obtain.

Weaknesses: Most of the methodology used was fairly straightforward, and has been used elsewhere before (to be fair, the authors did not claim they were novel)

The dataset had significant limitations (e.g. Yahoo was unfiltered, and twitter messages were post spam-filtering by twitter), thus comparing the two was problematic. URL blacklists were of limited use, so the authors tried to compensate with spamtraps. But those are bound to be incomplete. The heuristics used on top of that (whitelist Alexa1000 sites and filter for >1000 appearances in spamtraps) are fairly shaky at best.

The authors also can only show correlation but not causation in the evaluation of effectiveness. Without that, the conclusions about how to better fight spam based on these findings seem weak.

Comments to authors: There are some nice things about the paper:

1. it tries to do something that hasn’t been done before, comparing spam campaigns across email and OSNs
2. it does a good job of owning up to its limitations and being upfront about them.
3. it tries to ask some good questions, given the limitations of its data.

Unfortunately, the data just didn’t turn up much that was terribly interesting. The authors couldn’t do much about some of the limitations of the data itself. Those ended up reducing many conclusions to lower bound values and relative results (as opposed to absolute numbers). That took away most of the value of the paper.

Perhaps the most important question was on the issue of effectiveness. But the authors stopped digging after showing correlation. Without a clear cause, there really is not much one can do with the knowledge that spams simultaneously appearing on both platforms tends to be more effective.

Reviewer #4

Summary: There’s been a lot of work on spam detection, reputation, etc., but this work addresses cross-correlation, by looking at spam sent via e-mail and by twitter. The results are interesting, with sharp differences in overlap as well as click rates.

Strengths: The paper gathers a reasonably large data set and performs some interesting analysis on it, correlating between sources, and performing some analysis of click-through behavior.

Weaknesses: The analysis is not particularly deep - the paper is on the short side, with relatively large images and a fair bit of padding in the text. More analyses would definitely be helpful.

Comments to authors: I like this paper for the idea and the results, but do view it as somewhat preliminary. There’s analysis of overlap between domains, and some examination of the click-through behavior. In both cases, there are pretty interesting differences, and the paper has some conjecture, but little in the way of any deeper insight.

The paper also seems a little light in general. Section 3, for example, seems unnecessarily verbose for a short paper. It would be totally understandable to omit this level of description in a short paper, and given that there’s no “meat” in the paper until about the 4th page, it does drag a little. Similarly, the raw numbers in Table 1 are fine, but not really explored much, so it seems that it’s just padding. If there are interesting details in the table, they don’t really appear in the paper.

In the end, I think this paper represents a good start and a possibly-interesting area for further exploration. It’s got some interesting data and the start of an interesting analysis, but it’s not particularly deep.
**Reviewer #5**

**Summary:** This paper studies of behaviors of spammers across two different domains. Specifically, it compares spam URLs sent by over Twitter and emails. It finds significant correlation between the two, which suggests potential opportunities for designing spam filters that work across different domains and that leverage knowledge from one domain and apply it in another.

**Strengths:** The idea of conducting a cross-platform study of spam URLs is indeed novel, interesting and thought provoking. The finding of significant overlap in spam URLs tweeted over Twitter and email, especially for domains controlled by aggressive spammers, suggests potential opportunities for designing cross-platform spam filters.

**Weaknesses:**
- most of the results presented here are rather expected and not surprising or novel.
- the limitations of the data set cast a doubt over some of the findings.

**Comments to authors:** I really like the principal finding that there is significant overlap between spam URLs in Twitter and email, as a call for future work in designing cross-platform spam filters.

My biggest criticism of this work would be that beyond the high-level finding, most, if not all, of the remaining findings are rather expected and not surprising or insightful.

I wonder if you have sufficient data to conduct a temporal analysis of when spam URLs appear in Twitter and emails. That is, do they appear first in emails before they appear in Twitter or vice versa. If there is a considerable time difference between when spam URLs appear in Twitter and emails, it would suggest a clear advantage for cross-platform spam filters.

Finally, I hope this work would spur further research into designing cross-platform spam filters.

**Response from the Authors**

We first addressed the reviewers’ concerns regarding the claims of the paper. In each section, we added sentences or paragraphs describing how our results are affected by the limitations of the data set. We also reorganized parts of the introduction and conclusion to make them more aligned with the limitations of the data. We also expanded our comparison with the Monarch paper in the related work section. Second, we addressed the reviewer request for more detailed methodology: we added more explanations about the collection of the DNS data, setup and collection of spamtrap data, and extraction of the Yahoo data. Finally, we fixed several presentation issues: added percentages next to the absolute values in Table 1, corrected the typos in Figure 2, and added more explanation for why the numbers do not add up in Table 1.