Summary Review Documentation for

“Longtime Behavior of Harvesting Spam Bots”

Authors: O. Hohlfeld, T. Graf, F. Ciucu

Reviewer #1

Summary: The paper presents some very interesting data on the impact of honeypots to trap email-harvesting bots. This approach seems like a great idea, and the data size seems quite reasonable. The disappointing part is that the paper really does not dive into the interesting aspects of the data, but instead just gives a superficial view of the overall data without answering any real questions. This seems like a great opportunity wasted.

Strengths: Not much work has been done to empirically study how spam engines harvest email addresses. It’s an important problem that is relevant to many people, and the study has access to a dataset with significant potential. Some of the results are surprising and interesting.

Weaknesses: Instead of focusing on the 1-2 very interesting results and diving deeper, the paper chooses to take a rather shallow look across some very low hanging fruit. The result has some teasingly interesting results, but nothing substantial, and leaves the reader wanting.

Comments to authors: The data is interesting; the work is interesting, but the analysis should really go deeper to produce some interesting insights.

Some of the results are just superficial items that provide little insights/low impact. For examples: (1) use of anonymity services, (2) whether comment spammers harvest addresses, (3) geolocation analysis.

Instead, there are a couple of very interesting angles that needed much more analysis to provide potentially fascinating results. How do different types of address obfuscation protect users? This result is given so little attention, and yet is a critical question that could benefit the large majority of Internet users. How are harvesters using search engines? This seems like a potentially very interesting problem. Are there exploits/attacks at play? How is this being accomplished? Instead, the paper just punts and says this calls for further investigation. Why is that investigation not done here?

Reviewer #2

Summary: The paper uses a large dataset collected over 3 years to characterize email address harvesting (for spam) and comment spamming (e.g. in blogs). The authors characterize the usage of spamtrap addresses, the robustness of methods used to display email addresses on the web, the usage of anonymity services, and the correlation of address harvesting and comment spam. They highlight the observation that spammers use search engines to find email addresses. In addition, they make a number of observations that primarily confirm previous results using a newer and larger dataset. The analysis is quite extensive for a short paper and is based on an interesting measurement methodology using spamtraps.

Strengths: According to the authors, this is the first study that looks at comment spam (e.g., in blogs). The paper finds that there is no correlation between email and comments spammers. The two actions are handled by completely different bots.

The authors observe that spammers use search engines to find more emails.

The paper confirms a number of previous observations, like that anonymity services are rarely used for hiding the identity of spammers, using a larger and more recent dataset.

The paper is well written.

Weaknesses: Most of the results are not particularly novel or surprising. The results have a geographical bias (as with most studies of similar nature).

Comments to authors: What causes the knee in the solid line of Figure 4?

The results of Section 4.3 are interesting, but they seem to simply confirm previous results. It would be interesting to know how they compare with other studies and whether the authors see something new or not.

I liked the observations of Figure 2 and 3. But it was not clear what to take away.

The paper should provide a citation for the claim that the volume of spam has exceeded the volume of legitimate e-mail.

How many hosts were actually scanned?

In the x-axis label of Figure 4, it seems that ”(Days)” should be removed as the turnaround spread is in hours.

In Section 4.1, the paper indicates that 20% of the hosts were classified as search engines, while in Section 4.4 the paper mentions that 19% of the hosts submitted a user agent string of a major search engine. It seems, there is a small inconsistency here.

Reviewer #3

Summary: The authors analyze an extensive dataset of spamtrap email addresses seeded on web pages, where the addresses were encoded such that any subsequent use of them can be tied to a specific web page visit. The main findings are that often the visits came from search engines rather than individual systems, and attacks on the web pages to insert comment spam appear unassociated with the harvesting process.
In all this work is done fairly well. However, the topic is of only somewhat marginal interest these days, and their own data shows it’s of diminishing importance.

**Strengths:** Quite nice dataset and methodology. The analysis is generally thoughtful and investigates apt aspects of the data.

**Weaknesses:** Figure 2 shows a huge fall-off in harvesting, a fact that goes without comment. Given the use of other harvesting techniques such as bots trawling through infected systems looking for addresses (as discussed in [14], but not mentioned here), it appears likely that this study comes a few years too late to be of practical interest. The citation in support of it remaining “the primary mean[s] for spammers to obtain new target addresses” is from *2005*.

In addition to analyzing the harvesting process itself, the paper frames its other main contribution as being the investigation of whether comment spammers harvest e-mail addresses. However, the finding here is negative, which is unsurprising.

**Comments to authors:** Did you really have only 1,221 visiting hosts (or the subset considered “harvesters”) when you gave out about 15 million addresses? Did each host get on average 10,000 addresses? Something basic seems unclear here.

Why couldn’t you obtain DNS records for 10% of the hosts? Does this mean failed PTR lookups? If so, what about whois information, and/or checking them against blacklists?

For the unclassified comment spammers, what about manually analyzing a random subset?

How many Windows systems have port 3389 open even though they’re not servers? The success of the Morto worm suggests it’s lots.

I didn’t understand the comment ”A per-host based classification shifts the popularity of user agent[s] strings”.

The variation in figure 5 is quite striking, and makes me wonder whether just a few energetic hosts completely dominate the breakdowns.

**Reviewer #4**

**Summary:** The authors conducted a large scale experiment (3 years dataset) to study the early-stage behavior of spam campaigns, parimarily focusing on address harvesting and comment spam. They generate random addresses and embed these addresses into the web pages using six different presentation methods (e.g. mailto link, plain text, obfuscated email, etc.) to study robustness against harvesters. Also, they use HTML trap forms to study comment spam. The most interesting result is that in a few cases, the search engine webpage cache is used by harvesters to learn more e-mail addresses.

**Strengths:**

- This is a very nice study—one that I myself had thought of doing some time ago, but never actually did in earnest (I only planted a few tens of email addresses and then gave up). This paper takes the idea to completion, planting over 22 million email addresses.

- The paper’s findings that even simple obfuscation methods are still good enough to prevent spammers from harvesting email addresses is significant, and relevant for prevention of address harvesting.

- Although similar studies have been performed in the past, this is the first study to have been performed on such a large scale, and, due to the continually evolving nature of spam, it is worth re-evaluating techniques such as address harvesting every once in awhile.

**Weaknesses:** The paper could have been just as interesting (and perhaps more coherent) if it had just focused on address harvesting and ignored comment spam. The two problems are seemingly unrelated. It is interesting to note, nonetheless, that comment spammers do not appear to harvest email addresses.

**Comments to authors:**

- It seems that it is worth separating (1) the harvester; and (2) the spammer. I did not quite understand how the proposed technique can identify the harvester without access to the Web access logs. Are you using those as well? Or, is the address you generate somehow a function of the IP address that accesses the page, so that you can map it back to the IP address (and time) of a request? Otherwise, simply generating a new random email address for each access attempt doesn’t seem sufficient to identify the harvesters.

- In Figure 2, why does the number of harvester bots decrease? The paper acknowledges the decrease, but it does not explain what causes it. In general, I would expect the number of harvesters to remain somewhat stable. Something seems pretty off here (my understanding, the data, or both). Please explain better.

- The search engine finding is interesting, and it also suggests a possibly simple fix/defense to the harvesting problem: if a Web site recognizes the IP address or user agent of a search engine crawler is accessing the page, then it could obfuscate or simply not return email addresses.

- I was surprised to see in Figure 3 that the top ten countries of harvester bots are heavily skewed towards Germany. I’m suspicious because the study was performed in Germany. Might this be a function of the servers where the email addresses were hosted?

**Reviewer #5**

**Summary:** The authors conducted a large scale experiment (3 years dataset) to study the early-stage behavior of spam campaigns, parimarily focusing on address harvesting and comment spam. They generate random addresses and embed these addresses into the web pages using six different presentation methods (e.g. mailto link, plain text, obfuscated email, etc.) to study robustness against harvesters. Also, they use HTML trap forms to study comment spam. The most interesting result is that in a few cases, the search engine webpage cache is used by harvesters to learn more e-mail addresses.

- I liked the methodology used for data collection. However, for the most part, the authors present findings without any deep understanding into what is really going on.

**Strengths:**

- The paper’s findings that even simple obfuscation methods are still good enough to prevent spammers from harvesting email addresses is significant, and relevant for prevention of address harvesting.

**Weaknesses:** While the study of comment spam is interesting and new, it seems rather disjoint from the study of email spam. Overall
this paper presents some interesting results, but it is lacking deeper understanding into the insights behind the findings.

Comments to authors: Table 1: The percentage of spammed addresses is higher for random IDs than for realistic looking ones for all cases except for site G. Why is this the case? Are there bots that only harvest realistic-looking spamtraps? You may be able to answer this by studying the IP address or user-agent of the harvesters, and been able to tell if there is any bias towards IPs only harvesting realistic-looking spamtraps. Why is there one page that receives no spam? Any interesting insights here?

Figure 2: Why do harvester requests per month decreases in the last several months? I can imagine a few reasons, none of which the authors seem to describe or consider: i) Is it that if the bot visits the site one time, it does not come back or ii) is it that the sites were avoided by bots because bots identified the spamtraps.

Similarly for the comment spammers, why is the number of requests per month not decreasing? Is it that the same bots keep coming back to spam on the same sites?

Figure 3: Why is it the case that harvester bots have such a high bias to sites located in Germany? The authors show that the harvester bots are majority residential access networks. From the perspective of the botmaster, why would he/she be restricted to only use those bots to harvest sites in the same country? Could it be more a bias caused by the technique used to select the seed URLs to begin the crawl (maybe using Google Germany to get the seeds)?

Section 4.4: Those 5% hosts mimicking googlebot - where are they geo-located, are they DNS reverse-lookupable, can you tell whether they are residential ASes? The point is that it could be that the suspicious hosts are also Google servers, but hosted outside the Google AS.

Section 5: Given the finding about harvesters using search engine caches to optimize the harvesting process, I would expect a more detailed analysis into how is this actually done. Can you give more details into how they are accessing the google cache of webpages, is it through the "cached" link that comes on every Google search result? An alternative to this could be that malicious clients are using Google/Yahoo/Microsoft infrastructure to harvest. The authors already found some Amazon EC2 hosts performing comment spamming, so it is not that rare to find IP addresses in provider networks been used for suspicious activities.

End of section 4.4. It would have been nice to have more statistics to backup your claim that harvesting and spamming are done by the same entity. But currently you only mentioned an isolated case.

Are there any interesting characteristics of the spamtraps that received spam continuously for 3 years? Similarly, is there any interesting characteristic for the spamtraps that received spam after 3 years of being harvested?

Response from the Authors

To focus on the reviewers’ main interest in a detailed analysis of harvesting bots, we significantly reduced the comment spam discussion. Concretely, we only highlight the interesting aspect of whether address harvesting and comment spamming are simultaneous activities. We re-used the gained space for additional plots and discussions related to harvester bots.

Like the reviewers, we also question the declining trend in harvesting activities. While related reasons can be thought of, we decided not to include barely supported speculations.

The reviewers remarked on an interesting behavior in the usage period of our addresses, which is depicted in Figure 4 as a knee. A proper explanation of this behavior would require an extensive spam campaign analysis. Due to the complexity of such an additional and interesting study, we leave this for future work.

The reviewers requested discussion of how search engines are exploited by harvesters. While we agree on the value of such a discussion, our dataset was not designed for that particular purpose. In fact, one would need additional vantage points at the search engines, which are conceivably hard to install, in order to capture the interactions with the harvesters. An alternative way would be to reverse engineer the harvesting software referred to in the paper.