++

==========
PAM 2013 Review #14A
Updated Thursday 25 Oct 2012 1:27:45pm PDT

Paper #14: Searching for Spam: Detecting Fraudulent Accounts via Web Search
---

Overall merit: 3. Weak accept
Reviewer expertise: 3. Knowledgeable
Novelty: 3. Novel or a useful reappraisal
Technical merit: 3. Good/minor flaws
Scope: 3. Core Measurement paper
Presentation quality: 3. Mostly correct

===== Paper summary =====

The paper proposes to identify twitter spam by making web search on user accounts and marking as spam posts from users that have no presence on other social networks.

===== Comments for author =====

The idea itself is simple and to my knowledge new. A limitation of this idea is that, while it may catch an fraudulent account that was set up from the get-go as fraud but would not catch an account from a real user that was hacked. In fact, the proposed technique -- if adapted -- might actually encourage hacking of real users as a way to avoid detection. But that's a side note - it's always an arms race.
Then, the search of this name will result in a large number of hits. Beyond this, here are some comments:

* From your description, an even easier avoidance method would be for the spammer to simply select a very common word for display name (which I recall does not need to be unique). You describe an elaborate manual tuning for noise reduction, but I can't see which of the filters would eliminate all the search results for common words.

* I did not understand why you only made available tools but not the data set. You said at the end of sec. 4.1 that the data set would age quickly but that does not explain why you could not provide the snapshot that you used for this study. Since here no corporate data is involved, you are free to disclose your data, and I'd suggest that you do so.

* At the end of sec. 4.2., you speculate that some of your false positives come from accounts that used to be legitimate but were later hacked. My immediate question was why wouldn't you random-check this, until I found that you did at the end of sec. 4.3.

You badly need a forward pointer in sec. 4.2.

* In table 1, what's the diff between "display name" and "screen name"? Shouldn't be one of them "username"? Also, what about various combination of these filters?

* In the beginning of sec. 4.5, what's the "most frequently occurring domains"? Is it top-5 most frequent? Top-10? Overall, the tuning part seemed like a hack, with a lot of manual ad-hoc work. But in practice, while not pretty, it might be OK.

* At the end of sec. 4.5, you say that the blacklist stabilizes after 500 sets of results from fig. 3. How do we see this? On fig. 3, the point for 500 is far away from the point for 100, indicating that the FPR and TPR are quite different.

Poster: 3. Yes

==========

PAM 2013 Review #14B
Updated Thursday 1 Nov 2012 1:33:40am PDT

Paper #14: Searching for Spam: Detecting Fraudulent Accounts via Web Search

---

Overall merit: 4. Accept
Reviewer expertise: 2. Some familiarity
Novelty: 3. Novel or a useful reappraisal
Technical merit: 3. Good/minor flaws
Scope: 3. Core Measurement paper
Presentation quality: 3. Mostly correct

===== Paper summary =====

This paper presents a very lightweight yet surprisingly effective method of predicting whether a twitter message is spam: search the web for evidence that the tweeting party is a real person. (Hence it is possible to detect spam before it is even sent!) The paper shows that the technique achieves a true positive rate of over 74% and a false positive rate below 11%. But wait, you may ask, how is it possible to estimate true and false positive weights without some sort of ground truth? For ground truth, the paper looks at which accounts were suspended by twitter. It also makes a distinction between accounts that never sent any legitimate messages and accounts that appear to have been compromised and were then suspended. Of course twitter may not always be correct in making deciding which accounts to suspend, and the results might be interpreted as predicting which accounts twitter will ultimately suspend, rather than which accounts are actually spammers (still useful, s
ince it can be done up front), but the authors also manually inspect the accounts and messages sent on a smaller scale (hundreds) to confirm their results.

===== Comments for author =====

In the abstract "that require training and message content" is a bit confusing. How about "training and analyzing message content" ?

The paper should probably have more discussion about how spammers might go about trying to circumvent the technique described here if they knew it was being used. There is some suggestion that spammers wouldn't generally want to link their spam accounts across different social networks because the banning of one might lead to banning the others, but that's not wholly convincing. Perhaps by linking them, the odds that any of them get banned would become much lower? Also, if it all it takes is created a few more accounts across other services, perhaps the bar is not being raised that much. (Note, though, that the blackmarket price for a real-looking Facebook account is significant - the last quote I heard was several dollars.)

The discussion in Section 4 about the number of blacklists is a bit confusing. You are trying to filter out search query results for usernames and display names that would be present for any username or display name, whether or not they are spam accounts, right? E.g., if twitter makes every username available via a web-based directory, such a query result has no bearing on whether the account is spam. Maybe a few concrete examples would make this section more clear.

It's interesting that one of the authors of the paper has used the web scraping technique successfully in another context - collecting ground truth for a geolocation system.

Poster: 3. Yes

==========

PAM 2013 Review #14C
Updated Friday 2 Nov 2012 1:15:51pm PDT

Paper #14: Searching for Spam: Detecting Fraudulent Accounts via Web Search

---

Overall merit: 2. Weak reject
Reviewer expertise: 2. Some familiarity
Novelty: 3. Novel or a useful reappraisal
Technical merit: 2. Obvious flaws
Scope: 2. Borderline
Presentation quality: 3. Mostly correct

===== Paper summary =====

The paper suggest that one can use web search to determine if a twitter account is

likely to be an account created to send twitter SPAM.

===== Comments for author =====

As the basic idea seems reasonable the implementation details make many assumptions which are not well justified. Also the validation is overly optimistic. Here some more detailed concerns:

- People might not want to link there accounts for privacy reasons (e.g. see reaction of google linking accounts in the EU).
- Possible attack: search for user account and then create a twitter account with that account name. Not everybody has a twitter account so many account names in other services are available on twitter.
- small dataset in evaluation
- If the assumptions in 4.2 are correct then 40% of twitter accounts are fraudulent. That seams high and makes me question if the assumptions are incorrect which would mean the benefit of the algorithm are an overestimate.

Poster: 3. Yes