Consolidated Review of

Follow the Green: Growth and Dynamics in Twitter Follower Markets

1. Strengths:

The paper describes an approach to automatically detect "victims" in twitter follower markets. I found the paper is well written and easy to read. The twitter follower markets is an interesting topic. The detection mechanism proposed in this paper is also shown to perform well. Thorough analysis of Twitter follower markets along several dimensions.

2. Weaknesses

Evaluation with respect to twitter current behaviour is lacking. The paper built on several previous work of the same group of authors. This makes the technical novelty of this paper weak. Paper makes several assumptions (details below) that would benefit from validation. I have no idea how authors control the quality of accounts for the fair comparisons across markets. Authors (implicitly) assume that user accounts have similar quality across markets. Sometimes it could be true, but we cannot guarantee that.

3. Comments

Though the paper is well written and easy to read, I had to read and compare several authors' previous work in order to determine the new contribution of this work. It would be helpful is the paper clearly state the new contribution of this paper and difference from authors previous work. Also, as indicated below, at stages I think the presentation is difficult to follow and could be simplified and clarified.

The motivation of the work is clearly indicated and the contribution with respect to related work is also clarified. Although there seem to be other approaches to identify victims, in a way this approach is more automated and has diverse aims than other twitter follower studies. The work somehow is a follow up of some of the authors' work published in paper [28].

At stages I found the paper difficult to follow: in particular section 3 and 4 could be better written. I was particularly confused by the use of victim accounts in the first part and by 180 accounts registered in the second part. The relationship (if any) between these accounts and parts should be better clarified. I am also unclear of the difference between the terms market victim and market customer in this section. One more small point on section 3 is the use of Alr and Al as legitimate users: clearly this cannot be proven but as it becomes clear at later stage it is a good way of distinguishing statistically: perhaps I would use different terminology here.

At the end of section 4 it would be nice to indicate how quickly the accounts you found were shut by twitter, indicating if twitter already has a hang on the problem or not.

The technique used to detect automatically "anomalous" accounts is quite basic but effective. Using temporal changes in behaviour and followers is the obvious thing. The evaluation seems to show that it works. Overall it would be nice to know how much better this is with respect to what Twitter already does: can you shut accounts more quickly? You have a paragraph on this aspect at the end of section 6: I think a larger part of the analysis should be dedicated to this.

How big is the twitter follower market? What is the yearly revenue?

In section 4.2, the paper claims sheer volume of victim of twitter follower market. However, given twitter has over 500 million registered users, the number presented in Table 4 is about 0.1%. It does not sound very significant. It is not clear to me that victims you presented in Table 4 are all real user. What is the impact of the fact that these accounts are probably shut quickly on your analysis of bullet 2? Also, regarding bullet 4, you say you consider the account a victim if it tweets but I thought you started from accounts that are classified as victims right?

Fig 5 shows the trend of how market customers lose followers. How does this differ from normal users (or popular users) losing followers?

Fig 7-9 shows max increase in followers, consequent hours of followers decrease and consequent hours of constant followers. Using of max instead of medium makes the metrics more sensitive to outliers.

It is intuitive that market customers have distinct characteristics from normal users. However, the difference between market customers and popular users or big influencers who are not using follower markets. The detection method presented in this paper may detect both cases as market customers. The paper did not further analyze the false positives. It would make the paper more interesting if this aspect can be further explored.

1. In inferring market customers, starting with victims, you find their friends that have not (in essence) posted a similar tweet. You also mention that (in many cases) victims are created when they add certain OAuth apps or share their account credentials with a website. Thus there is a point in time before which the account was not a victim, and a point in time after which the account was a victim. However, you do not appear to ignore friends that the account had before it became a victim (not sure if that is possible given your dataset). But in essence, say a user friends Celebrity X before he was a victim, and subsequently gets victimized, then your algorithm will label Celebrity X as a customer, is that correct? If so, your customer numbers may be inflated. It would be good to quantify any such inflation to the extent you can.
2. Re: victims unfollowing customers, do you have an hypothesis on why victims persist to remove friend links, but do not unauthorize the OAuth app that is controlling their account?

3. Further to the point of victims unfollowing customers (presumably based on tweets made by customers that are less engaging that tweets made by legitimate twitter users), have you validated that it is the lower engagement of tweets that cause unfollowing behavior? E.g., if you acquire some victims, and then retweet tweets only form engaging customers (e.g., Obama etc.) do your victims have longer retention time?

4. The number "30% of legitimate users did not experience changes in their followers" may be biased by the fact that A_lr have far fewer followers than A_c. What fraction of A_lr has fewer than 100 users? A follower count change of 10 followers out of a million corresponds to a change of 0.001 followers for an account with 100 followers. For a fair comparison, a normalized change in followers makes more sense. What fraction of A_c experienced a 1% change in follower count? Vs. what fraction of A_lr experienced a 1% change?

5. In Section 6.2 you suggest several filters. How would the adversary react? E.g., an adversary could easy avoid the "more followers than friends" and "followers to friends ratio" filters.

6. Have you presented A_c to Twitter? Do you believe the evidence you gathered is sufficient to move Twitter to ban members of A_c? E.g., if Twitter requires evidence of money exchanged (I don't know if it does), your method cannot provide that evidence. Or if there may be false-positives in A_c, Twitter cannot justify banning all of A_c. Is there a way to characterize false-positives in your result?

4. Summary from PC Discussion

The paper describes an approach to automatically detect "victims" in twitter follower markets.

Strengths:

- Well written.
- Nice analysis of twitter followers market.

Weaknesses:

- Some limitation in comparisons between markets (assumption of similar quality) and other assumptions not validated
- Comparison with current twitter technique not discussed

5. Authors' Response

We thank the anonymous reviewers for their comments. In the following, we summarize how we addressed the concerns that were raised.

First of all, the detection that we propose in the paper deals with detecting market customers, and not victims. Currently, the only countermeasure that Twitter has in place against follower markets is blocking victim accounts, or shutting down the offending OAuth applications that are used by the market operators to control their victims. These countermeasures do not scale, because market operators can easily create new OAuth applications, or recruit new victims. Instead, detecting and blocking market customers is more effective, because it hurts the market's business model. As Twitter is currently not blocking accounts for the mere fact of being market customers, we cannot compare our detection mechanism with established techniques. If a customer account is blocked by Twitter, it is because the account posted messages that were considered spam. Because, not all market customers post spam messages a more focused approach to detect market customers is warranted.

Comparison with previous work: Our WOSN paper introduced the concept of Twitter follower markets, and performed an initial analysis of the phenomenon. In this paper, we perform a comprehensive analysis, by subscribing our own victim accounts, buying followers from these services, and studying the follower dynamics of a multitude of accounts that are involved in Twitter follower markets. In addition, we develop and evaluate a technique to detect market customers. Twitter could leverage this technique to shut down accounts that bought followers, since this practice violates their terms of service. We clarified the contributions in the Introduction section to reflect this observation.

The size of Twitter follower markets: Accurately estimating the size of Twitter follower markets is challenging. In the paper, we identified a significant number of victims and customers. However, since our view of the social network is partial, we are probably missing a large number of accounts that are involved in follower markets. Similarly, it is hard to evaluate false positives. We cannot have a definite answer on whether a Twitter account bought followers or not. However, because of the way in which we established ground truth, we are confident that our technique is reliable in detecting customer accounts.

Identifying market victims: The accounts that we used for the experiment were freshly created, and had no legitimate friends. Therefore, the accounts that they started following were either market victims or market followers. We clarified this observation in the camera-ready version of the paper.

We can only speculate why victims do not revoke authorization from malicious OAuth apps. A reason could be that victims want to gain more followers, and only selectively unfollow particularly annoying accounts.

We decided to use absolute numbers in our analysis, rather than fractions, because follower markets sell a fixed number of followers (for example, 3,000). By looking at absolute numbers, we can observe increases in followers that are indicative of a purchase of followers.

We added a discussion section about how an adversary could react to our detection in the camera-ready version of the paper: In Section 6.1, we discuss possible evasions against the dynamic classifier, while in Section 6.2 we discuss how an attacker could evade detection by the static methods. We did not present our results to Twitter, because Twitter does not have a system in place to detect and block market customers. However, should Twitter be interested, we would gladly collaborate to establish such a system based on the approach presented in our paper.

As a last remark, our work does not assume that the different markets have a similar quality. In Section 5.3, we show that some markets provide followers of a lower quality, which are often suspended. Instead, our work aims at identifying characteristics that are typical of market customers, and common among different markets.