Followers or Phantoms?
An Anatomy of Purchased Twitter Followers

Anupama Aggarwal
IIIT-Delhi
anupamaa@iiitd.ac.in

Ponnurangam Kumaraguru
IIIT-Delhi
pk@iiitd.ac.in

ABSTRACT
A significant percentage of OSN users utilize various methods to drive and manage their reputation on OSN. This has given rise to underground markets which buy/sell fraudulent accounts, ‘likes’, ‘comments’ (Facebook, Instagram) and ‘followers’ (Twitter) to artificially boost their social reputation. In this study, we present an anatomy of purchased followers on Twitter and their behaviour. We illustrate in detail the profile characteristics, content sharing and behavioural patterns of purchased follower accounts. Our study highlights the key identifiers for suspicious follow behaviour. We then built a supervised learning mechanism to predict suspicious follower behaviour with 89.2% accuracy.

1. INTRODUCTION
OSN like Twitter, Facebook and Instagram are being used by Internet users to interact and spread information by enabling them to maintain their online identity. This online identity is based on content sharing and interaction patterns. To boost the reputation and popularity of their online social profiles, users utilize various methods like sharing interesting content, attracting more ‘likes’ and ‘followers’. This has led to the creation of an underground fraudulent market which promises to boost the reputation of online social profiles by selling ‘likes’, ‘comments’ and ‘followers’. Overall, Twitter follower markets provide two popular purchasing schemes - Freemium and Premium. In freemium scheme, the customers do not have to pay any money to gain followers, but only to authorize merchant’s Twitter app. However, in return, the customer becomes part of the phony follower network. In case of premium scheme, the customer has to pay money to the merchant to gain followers. This study has the following research contributions:

- We present an anatomy of the purchased Twitter followers. We characterize the profile attributes and the behavioural features of the purchased followers. We also compare their characteristics with legitimate users.
- We identify key indicators to distinguish between suspicious following behaviour from that of genuine Twitter users. We use these identifiers and built a supervised learning mechanism which identifies suspicious following behaviour with an accuracy of 89.2%.

2. ANATOMY OF PHONY FOLLOWERS
We find key identifiers which can be helpful to distinguish from legitimate users and hence help us to build an effective detection model.

Low social engagement.
We observed that a large fraction of purchased accounts post only retweets instead of original content. We further explore whether these users retweet the content of their friends or not. If $RT_{count,i}$ is the number of tweets the user has retweeted of her friend $i$, and she has $N$ friends, then

$$RetweetRatio = \frac{\sum_{i=1}^{N} RT_{count,i}}{N \times RT_{total}}$$

$RT_{total}$ is the total number of retweets done by the user. This Retweet Ratio quantifies the number of friends a user has retweeted and the number of times she retweeted them. Similarly, we define @-mention ratio to determine whether the user engages in conversations with her friends and to what extent. We observe that the highest Retweet Ratio score is 0.45 and the @-mention ratio is 0.35. This shows that though a large fraction of purchased accounts post only retweets, its not the tweets of their friends which they are retweeting. Similarly, low @-mention ratio suggests that purchased followers do not mention their friends. We found the maximum @-mention ratio with the followers of purchased users to be 0.32. This indicates that purchased followers are low quality users and do not engage in conversations with their friends or followers.

Low social reputation.
We now look at the relationship between amount of followers and friends for purchased follower accounts. On Twitter, ‘followers’ of a person are the users who subscribe to the posts of that person, i.e., who ‘follow’ her. The ‘friends’ of a person are the users whom she subscribes to. The average number of followers per existing account is 68 and the average number of friends is 60 on Twitter.

We observe that 94% purchased followers have the follower/friends ratio as only 0.1 and none of the purchased followers had more followers than friends. Low follower/friends ratio indicates that the user does not have a good following, therefore indicating a low social importance.

To measure the social influence, we use Klout score. ‘Klout’ is a popular tool to measure influence based on various factors like followers, friends, retweets and favourites. The average Klout score for the social media users is 40. However, we found that 90% of the purchased followers had a Klout score of less than 20. This shows that these accounts do not involve in discussions with other users and have a low influence score.

1 http://support.klout.com/customer/portal/articles/679109-what-is-the-average-klout-score
High unfollow entropy.

We found that the purchased follower unfollowed a large number of users regularly. To quantify this behaviour, we calculated the unfollow entropy of all the purchased followers. We observed each purchased follower over a span of 15 days and collected her hourly followers. We define normalized unfollower entropy $H_u$ for a user $u_n$ as the following:

$$H_u = \frac{\sum_{i=1}^{T} p_n(f_i) \log(p_n(f_i))}{N}$$

where, $p_n(f_i)$ is the probability that the user $u_n$ will unfollow at time $t_i$. The probability function is defined as

$$p_n(f_i) = \frac{ucount_i}{\sum_{i=1}^{T} ucount_i}$$

where $T$ is the number of days for which we monitor the purchased follower and $ucount_i$ is the number of users she unfollowed on $i^{th}$ day. A higher value of unfollow entropy signifies that the user exhibits a suspicious unfollow pattern.

Figure 1 shows that a large fraction of purchased followers have a high unfollow entropy. The normalized entropy rate for 23% purchased followers is as high as 0.76 and only 8% users have a normalized unfollow entropy less than 0.21.

![Figure 1: Unfollow Entropy Rate for Purchased Followers. A large number of followers in our data follow-unfollow their friends multiple times.](image)

3. PREDICTION OF PHONY FOLLOWERS

Now, we build a supervised predictive model to detect suspicious following behaviour on Twitter.

Features for Classification.

For our prediction task to detect suspicious following behaviour, we explore user profile, network, content and user behaviour based features. Table 1 enlists all the feature sets we used for our prediction task.

Experimental Setup and Classification.

For our classification experiment, we consider the 170k public purchased followers as our true positive dataset of suspicious follow behaviour. For the negative class (legitimate follow behaviour), we pick random 170k users from Twitter stream using the streaming API. We use a non-linear SVM with the Radial Basis Function (RBF) kernel for our experiment.

Classification Results and Evaluation.

Table 2 shows the confusion matrix for our classification task. The confusion matrix defines the percentage of false negatives and false positives. We were able to accurately classify 88.5% users with suspicious follow behaviour and 89.9% users with legitimate behaviour. This shows that we are able to detect suspicious following behaviour to a good extent.

![Table 1: Description of the feature sets used for prediction of users with suspicious following behaviour.](image)

<table>
<thead>
<tr>
<th>Set</th>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>A User Profile</td>
<td>presence of bio</td>
<td>presence of URL in bio</td>
</tr>
<tr>
<td></td>
<td></td>
<td>number of posts</td>
</tr>
<tr>
<td>B Network</td>
<td>follower / friends ratio</td>
<td>number of followers</td>
</tr>
<tr>
<td>C Content</td>
<td>hashtags per tweets</td>
<td>span words used per tweet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>length of tweet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>number of languages used</td>
</tr>
<tr>
<td>D Behaviour</td>
<td>number of RTs per tweet</td>
<td>number of RTs per tweet</td>
</tr>
<tr>
<td></td>
<td></td>
<td>@mentions per tweet</td>
</tr>
</tbody>
</table>

![Table 2: Confusion Matrix – Classification Results of distinguishing legitimate users from those exhibiting suspicious following behaviour.](image)

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Suspicious</th>
<th>Legitimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>88.5</td>
<td>11.5</td>
</tr>
<tr>
<td>Suspicious</td>
<td>9.7</td>
<td>89.9</td>
</tr>
</tbody>
</table>

Feature Importance.

We found that behavioural features are important to detect suspicious behaviour. Unfollow entropy rate plays an important role; it is defined as the frequency with which the user is unfollowing her friends over time. Some of the most informative features we received after our classification task were unfollow entropy, RT-engagement ratio, @mention-engagement ratio, Language Overlap and Social Reputation. The other informative and discriminative features were the use of multiple hashtags and spam words in the tweets.

4. CONCLUSION

In this study we explored the dynamics of purchased follower accounts. We found some characteristic features of users which exhibit suspicious follow behaviour. We investigated the behavioural features of the followers purchased from underground Twitter follower market and found that a large fraction of users feep following and unfollowing their friends at regular basis - an activity which is unusual for a legitimate account holder. We thus define the term unfollow entropy to measure the rate of unfollow over time. In order to understand the dynamics of purchased follower accounts, we divided our study into two parts. In the first part, we studied the properties of users with suspicious follow activity and how they are different from regular Twitter users. In the next part, based on the discriminative features, we used supervised learning methodology to detect suspicious follow behaviour from regular behaviour. We received an overall accuracy of 89.2%.