Follow the Green: Growth and Dynamics in Twitter Follower Markets

Gianluca Stringhini, Gang Wang, Manuel Egele†, Christopher Kruegel,
Giovanni Vigna, Haitao Zheng, Ben Y. Zhao
Department of Computer Science, UC Santa Barbara
†Carnegie Mellon University
{gianluca, gangw, chris, vigna, htzheng, ravenben}@cs.ucsb.edu, megele@cmu.edu

ABSTRACT

The users of microblogging services, such as Twitter, use the count of followers of an account as a measure of its reputation or influence. For those unwilling or unable to attract followers naturally, a growing industry of "Twitter follower markets" provides followers for sale. Some markets use fake accounts to boost the follower count of their customers, while others rely on a pyramid scheme to turn non-paying customers into followers for each other, and into followers for paying customers. In this paper, we present a detailed study of Twitter follower markets, report in detail on both the static and dynamic properties of customers of these markets, and develop and evaluate multiple techniques for detecting these activities. We show that our detection system is robust and reliable, and can detect a significant number of customers in the wild.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences; K.6 [**Management of Computing and Information Systems**]: Security and Protection

General Terms

Measurement, Security

Keywords

Twitter, Follower Markets, Sybils, Online Social Networks

1. INTRODUCTION

Microblogging services such as Twitter have become important tools for personal communication as well as spreading news. Twitter users can "follow" accounts that they find interesting, and start receiving status updates that these accounts share in real-time. As Twitter use grows, the influence and reputation of a person or business entity are increasingly associated with their number of Twitter followers [8, 18]. Third party services, such as Klout, estimate the influence of accounts ranging from normal users to celebrities and politicians [19] based on a series of features such as the number of followers and the frequency with which content is re-shared.

IMC'13, October 23–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2504730.2504731.

For many users hoping to achieve fame and "influence" on Twitter, growing a significant follower population is a difficult and time consuming process. Some follow random users in the hope that these users might follow back. Others join groups where each member agrees to follow all the others in the group. Unfortunately, none of these techniques are efficient enough for users who need large numbers of followers fast [16].

The demand of quick Twitter followers has led to the growth of an industry that caters to users who want to quickly grow their population of followers, even if it means paying for them. Some of these "customers," such as politicians or celebrities, might want to give the appearance of a large fan base [4, 9, 10]. Other, more malicious entities seek followers so they can quickly spread malware and spam [12, 14, 29, 30]. We refer to such enterprises as *Twitter follower markets* and to the operators as *follower merchants*.

There are two ways for follower merchants to deliver followers to their customers. One way is to create fake accounts that imitate real users [10, 31, 33, 34]. However, these accounts have less "value," as quantified by popular metrics that establish reputation based on quantities such as a user's follower to followee ratio and tweets per follower. Fake accounts, or *Sybils*, provide lower additive value in this system, given their small follower counts and low volume of tweets. On the other hand, legitimate accounts are more attractive as followers to cybercriminals, because these accounts have real followers, share content, and add real value to customers seeking a higher reputation. Unfortunately for these markets, legitimate users are unlikely to voluntarily follow their customers, unless they were first compromised or controlled without their knowledge[1].

In our work, we perform a detailed measurement study of Twitter follower markets. Note that unlike prior work that studied black markets for social network accounts [25, 31], these merchants only provide "followers," not access to compromised or fake accounts. In our work, we observe two types of behavior in Twitter follower merchants. First, some merchants provide their customers with fake accounts (*i.e.*, Sybils). Second, we also observe the rise of a new type of follower merchants centered on pyramid schemes. In this case, follower merchants lure in unsuspecting users with promises of free followers, compromise their accounts, and then add them as new followers to each other and (more importantly) to other paying customers; Since these accounts are compromised by the market operators, we call them *victims*. We call the operators of pyramid markets *pyramid merchants*.

In this paper, we perform a comprehensive study of Twitter follower markets. First, we study in detail both the static properties

[1]We consider an account as compromised if a third party obtained access to it, and is using it in a way that violates Twitter's terms of service [3].

and the dynamic behavior of market customers. Then, we develop and evaluate multiple techniques for detecting these activities. We first introduced the notion of Twitter follower markets and some ad hoc observations of pyramid merchants in our preliminary work [28]. Compared to our preliminary work, we make the following contributions:

- We explored methods to detect follower markets and proactively obtained ground-truth composed of 69,222 victim accounts, and 2,909 market customers.
- We analyzed the characteristics of victim accounts and market customers, with particular focus on the dynamics of their followers. Intuitively, we expect customers of follower markets to experience dramatic increases in the number of followers followed by a steady decrease as compromised accounts unfollow. This happens because victim accounts did not willingly follow the customers, and often times they find the content that the customer posts uninteresting.
- We developed a detection system that uses follower dynamics to detect the customers of follower markets. We show that our system is robust and reliable, and that it is able to detect a large number of customers in the wild.

2. METHODOLOGY AND BACKGROUND

According to anonymous follower merchants, follower markets have already become a multimillion-dollar business [26]. Driven by their increasing popularity and effects on Twitter's ecosystem, we want to understand via empirical measurements the key characteristics of these markets, particularly the newly-emerged *pyramid merchants*, and develop effective countermeasures against them. Our study is motivated by two observations.

First, there is no concrete understanding on the characteristics of these markets as well as the key factors leading to their success and impact. To the best of our knowledge, the only known prior work is our preliminary study which detected a follower market [11].

Second, there are no effective countermeasures against these markets and follower merchants. Existing detection techniques can spot fake accounts or spammers [5, 14, 29], but are not effective against pyramid merchants as the purchased followers are real users. Instead, we believe that a fundamentally different method can effectively address this challenge. Specifically, if one can identify customers who bought Twitter followers from these markets, then once Twitter starts to systematically and heavily penalize or even suspend their customers, the markets will soon lose their revenue stream and eventually go out of business.

In the remainder of the paper, we describe our efforts to empirically characterize the accounts involved in follower markets and build detection algorithms. In Section 2.1 we briefly introduce the operations of Pyramid merchants to provide background for our study. We then discuss our data collection process (Section 3), followed by analysis on the market characteristics (Section 4) and its key players (Section 5). We then use insights from our analysis to develop a comprehensive detection system for identifying the market customers and to experiment on real Twitter data (Section 6).

2.1 Background: Pyramid Merchants

Unlike typical Twitter follower markets that trade fake accounts, pyramid markets sell real accounts as followers. They usually offer two kinds of subscriptions: *premium* and *free*. Premium users pay to get the services from the market, and are mostly companies or

Figure 1: A Twitter follower market website. The market provides two options, depending on how fast the customer wants to obtain his followers.

Figure 2: A tweet advertising a Twitter follower market. The link in the tweet points to the homepage of the Twitter follower market being advertised.

individuals who are interested in increasing their follower count and reaching a broader audience (we refer to them as *customers*).

Free users, on the other hand, are typically offered a small number of followers for free. In exchange for this service, free users are requested to give control of their account to the pyramid merchant. The merchant gains control of a user account by asking the user to authorize an OAuth application [2], or by having the user give away her profile name and password. Because these "free" subscribers give control of their account to a third party, we consider them as compromised, and refer to them as *victims*. The pyramid merchants leverage their victims to carry out many tasks, from following other accounts to periodically tweeting advertisements for the markets (see Figure 2 for an example). Although certain pyramid markets announce to their victims that their accounts might be used to follow other people or to send tweets, this practice violates Twitter's term of service [3].

In their attempt to shut down market operations, Twitter blocks the OAuth applications that are used by such schemes. However, pyramid merchants overcome this problem by periodically creating new OAuth applications and using the victims' credentials to authorize such applications [11]. Furthermore, to hide their involvement in any follower market, customers who purchased followers typically add these followers slowly. In fact, some follower markets advertise that it can take up to one month for a customer to add 3,000 followers.

3. DATA COLLECTION

In this section, we present our efforts to collect a large set of Twitter accounts that actively interact with follower markets, essentially building the ground-truth data on customers who bought followers from the markets and victims who were compromised by the markets and traded as followers. Our data collection includes two steps. First, we locate popular follower markets. Second, we collect accounts of customers and victims as well as legitimate Twitter users for further analysis. In this section, we describe these two steps in detail.

3.1 Locating Follower Markets

In general, a Twitter follower market offers services through a website where customers can directly make purchases. An example is shown in Figure 1. To locate as many markets as possible, we explored three different approaches, including *locating suspicious clusters of Twitter accounts*, *searching for advertisement tweets*, and *querying a search engine*. We found that the search engine-based approach is the most effective and thus used this technique for our subsequent data collection. Next, we discuss each method in detail.

Clustering Twitter Accounts. We first encountered evidence of Twitter follower markets from our previous study [11], in which we detected a viral Twitter campaign advertising a follower market (`Bigfollow.net`). To attract customers, follower markets often send advertising tweets (with a link to the follower market's website) using compromised Twitter accounts under their control (*i.e.*, victims). Therefore, one way of detecting these markets is by grouping Twitter accounts that demonstrate similar suspicious behaviors (*e.g.*, sending similar tweets) [11]. This method, however, has two key limitations. First, not all Twitter follower markets use Twitter to promote themselves (in particular, the non-pyramid markets typically do not); therefore this method cannot provide full coverage of popular markets. Second, this method also requires significant manual effort to examine suspicious clusters and exclude benign ones and those representing non-Twitter-market campaigns.

Searching Advertisement Tweets. The second approach is to directly search for tweets advertising the follower markets using the Twitter API. These tweets usually contain links pointing to the markets and keywords like "buy Twitter followers" or "get more followers" (Figure 2). Therefore, by searching for specific keywords in the tweet stream, we could identify various markets. After performing the above keyword search from the 1.4 billion tweets that we collected from Twitter's stream, we identified eleven Twitter follower markets that send advertisement tweets in bulk. The limitation of this approach is that it requires access to a large volume of tweets and it is also prone to false positives. Since our tweet dataset was collected between May 13, 2011 and August 12, 2011, we can only identify markets that actively advertised during this time frame.

Querying Search Engine. As it is difficult to gain a comprehensive picture of the follower markets directly from Twitter, we decided to leverage search engines. To attract customers, these markets should be able to make their sites indexed by Google. We confirmed this hypothesis by sending the query "more Twitter followers" to Google and analyzing the returned results. Other than links to the actual Twitter follower markets, we also found links pointing to news or blog articles talking about Twitter follower markets. For this reason, we cannot solely rely on the results returned by Google to detect follower markets.

To distinguish between real Twitter follower markets and benign websites, we developed a Support Vector Machine (SVM) classifier. We extracted 30 features from the URL, domain, and HTML content of the market website. The detailed features are listed in Table 1. The keywords were chosen by mining the web pages of known Twitter follower markets. This effectively exclude news article pages talking about markets, since the index page of news portals usually has no Twitter follower markets features.

Feature Category	Feature Description
URL (3)	Level of URL depth Length of URL Length of domain
HTML (7)	Number of $ sign Number of hyper links Number of outgoing links Number of image tags Number of Javascript tags Number of buttons Word count
Keywords (20)	twitter, followers, buy, social, service, facebook, services, youtube, views, likes, contact, order, within, fans, real, marketing, 100, privacy, account, website

Table 1: Features to locate Twitter follower markets by querying search engines.

Market	$ for 10K Followers	$ for 1K (Re)Tweets	Twitter Only?	Pyramid?
newfollow.info	$216	N/A	YES	YES
bigfolo.com	$91.99	N/A	YES	YES
bigfollow.net	$70	$500	YES	YES
intertwitter.com	$65	$550	NO	NO
justfollowers.in	$95	N/A	NO	YES
twiends.com	$169	N/A	YES	YES
socialwombat.com	$49	$79	YES	NO
devumi.com	$64	$97	NO	NO
hitfollow.info	$214	N/A	YES	YES
plusfollower.info	$214	N/A	YES	YES
buyactivefans.com	$40	N/A	NO	NO
getmorefollows.com	$127	N/A	YES	NO
followmania.co	$48	N/A	YES	YES

Table 2: Popular Twitter follower markets. Different markets have different prices, and some of them are not Twitter-specific. The follower markets that do not use a pyramid scheme are likely selling fake accounts as followers.

To evaluate the effectiveness of this method, we manually verified the first 100 websites returned by Google, and constructed a ground-truth set with 56 Twitter follower markets and 44 benign websites. A ten-fold cross-validation leads to a 97% accuracy. On February 2013, the query "more Twitter followers" offered 7,210,000 results in Google, while Google only returned the first 680 sites. We ran the classifier on all the returned 680 websites and located 303 Twitter markets. In Table 2 we list the 13 highest-ranked markets by Google. We believe this approach provides a reasonable coverage of current popular Twitter follower markets.

3.2 Locating Customers and Victims

Our next step is to locate a large set of accounts corresponding to customers and victims of the follower markets, as well as legitimate Twitter users. We use these sets of accounts as ground-truth for our further experiments.

Market	Victims	Customers
BigFollow	16,185	2,781
Bigfolo	1,404	37
JustFollowers	20,897	91
NewFollow	10,307	N/A
InterTwitter	20,429	N/A
Total	69,222	2,909

Table 3: Identified market victims and customers among the six Twitter follower markets that we monitored.

Collecting Market Victims. To identify a large number of ground-truth victim accounts, we purchased followers from the most popular follower markets in Table 2: *NewFollow*, *Bigfolo*, *InterTwitter* and *JustFollowers*. Here *InterTwitter* is the only non-pyramid market as it does not offer free-subscriptions.

Specifically, we registered one account as premium user for *Bigfolo* and *NewFollow*, and two accounts each at *InterTwitter* and *JustFollowers*. Since our accounts were newly-created with no followers, we consider any account that started following them a market victim. We made our purchases on March 27, 2013. Within two weeks we identified 53,037 victims (1,404 on *Bigfolo*, 20,429 on *InterTwitter*, 20,897 on *JustFollowers*, and 10,307 on *NewFollow*). However, at the end of the first week, Twitter banned two of our premium accounts, one at *JustFollowers* and another one at *Bigfolo*.

Unfortunately, we were not able to buy followers from the market *BigFollow*, the third highest ranked market, because their website did not accept payments from our credit card. Fortunately, the victims in this market are very recognizable, because they post very specific tweets (i.e., the tweets contain the hashtag *#BigFollow*). Therefore, we searched Twitter for tweets advertising this market, and considered any account posting such tweets as a victim. In total, we identified 16,185 victims for this market.

Overall, we were able to identify 69,222 victims. We refer to them as $\mathbf{A}_v$. A detailed breakdown of victims per market can be found in Table 3.

Collecting Market Customers. The core part of a Twitter follower market business is to sell followers. Therefore, to understand the phenomenon we need to monitor and study the characteristics of a set of customers in the wild. The problem with selecting a set of customers is that it is hard to determine which accounts purchased followers on Twitter. To overcome this problem, we registered 180 newly-created Twitter accounts as victims to the target follower markets (*BigFollow*, *NewFollow*, *Bigfolo*, and *JustFollowers*). We call this set of accounts $\mathbf{A}_g$. Since the accounts in $\mathbf{A}_g$ were newly-created, and had no followers or friends [2], we can assume that any account that established a friend or follower relation with them is somehow involved in the follower market (as either a victim or a customer). We did not subscribe to *InterTwitter*, because their website does not offer a free subscription service. This in fact raises an interesting question: where do the victims of this market come from? One possibility is that *InterTwitter* does not use free-subscribers as victims but uses massively-created fake accounts to follow their customers. We will further explore this intuition in the data analysis section (Section 5).

As previously explained, after signing up to the follower market, victims will start following customers, as well as other victims. For this reason, we cannot build a set of known customers just by looking at the friends of the accounts in $\mathbf{A}_g$. Instead, we leverage the following observation: market victims periodically post tweets that advertise the market. Such tweets are likely to be similar across the victims of the same market. To detect customers among the friends of the accounts in $\mathbf{A}_g$, we apply the following algorithm. This process returns a set $\mathbf{A}_c$ of identified customers.

1. At the beginning, the set of known customers $\mathbf{A}_c$ is empty.
2. For each account a in $\mathbf{A}_g$, we retrieve the tweets posted by a. We call this set $\mathbf{T}_a$. We also retrieve the set of friends that a has, and call it $\mathbf{F}_a$.
3. For each account b in $\mathbf{F}_a$, we extract the set of tweets posted by b. We call this set $\mathbf{T}_b$.
4. For each tuple $< t_1, t_2 >$ composed of one tweet from $\mathbf{T}_a$ and one tweet from $\mathbf{T}_b$, we compare t_1 and t_2. If t_1 and t_2 are similar, we consider b as a victim, and move to the next account in $\mathbf{F}_a$. More precisely, we consider t_1 and t_2 to be similar if they share four or more consecutive words (4-grams). This similarity metric has already been used in previous work, and has proven to be robust [11]. Note that the accounts in $\mathbf{A}_g$ never posted any legitimate tweet, therefore our approach will only match tweets that advertise follower markets and not, for example, popular tweets that both accounts happen to have shared.
5. If no pair $< t_1, t_2 >$ resulted similar, it means that b has never advertised the follower market. Therefore, we assume that the owner of b paid to get her followers, and add b to the set of known customers $\mathbf{A}_c$.

In total, we identified 2,909 market customers by using the described algorithm, 2,781 from *BigFollow*, 37 from *Bigfolo* and 91 from *JustFollowers*. We did not manage to identify any customer for the market *NewFollow*, because Twitter suspended our victim accounts shortly after we subscribed them, probably because the tweets advertising the market that these accounts sent were considered spam. A summary of identified customer and victim accounts is shown in Table 3. We noticed a disparity in the number of customers identified from these markets. The reason might be that some markets are more successful than others in attracting customers; alternatively, it might be that some markets have more victims, and therefore each victim follows a smaller number of customers. We will investigate these possibilities in Section 5.4. In the rest of the paper, we refer to the entire set of 2,909 customers as $\mathbf{A}_c$.

Selecting Legitimate Users. Finally, we wanted to gather a large number of regular Twitter users to draw comparison with market customers. At a high level, we selected two sets of legitimate users for different purposes. First, we picked one set of 2 million randomly-sampled users $\mathbf{A}_{lr}$ from the general Twitter population, which serves as baseline for our profile analysis. The second set, that we call $\mathbf{A}_l$, is our key legitimate user dataset, which focused on legitimate users who are comparable with market customers. For $\mathbf{A}_l$, we excluded users that are obviously non-customers (*i.e.* users with a relatively low follower count). To this end, we constructed $\mathbf{A}_l$ by extracting another two million unique accounts whose number of followers exceeded 100. In our later analysis (Section 5), we will show that choosing 100 as a threshold is reasonable. Both legitimate user sets were sampled from our dataset of a stream of 10% of all public Twitter messages [11]. Note that we cannot be sure that the accounts in $\mathbf{A}_l$ and $\mathbf{A}_{lr}$ did not purchase followers on Twitter. However, since the accounts were collected at random, we are confident that they are representative enough of the Twitter

[2]In Twitter's jargon, a friend is a profile that the account follows.

Market	Tweets	Victims
BigFollow	662,858	90,083
Bigfolo	4,732,016	611,825
JustFollowers	302	257
NewFollow	77,865	38,341
InterTwitter	0	0
Total	5,473,041	740,506

Table 4: Tweets advertising markets and victims identified in the wild. These results show that some markets are significantly larger than others.

population, and that they will show different characteristics than the set $\mathbf{A}_c$, that is solely composed of customers. In fact, the results reported in the rest of the paper confirm this assumption.

Summary. In total, we identified 4 million legitimate users ($\mathbf{A}_{lr}$ and $\mathbf{A}_l$), 69,220 market victims ($\mathbf{A}_v$) and 2,909 market customers ($\mathbf{A}_c$) for analysis. Using this dataset as our ground-truth, we analyze the behavior patterns of different parties that interact with Twitter follower markets. Then, we leverage potential behavior features to build systems to detect customers on Twitter. We will illustrate the details of the performed analysis and of the proposed detection in the following sections.

4. ANALYZING FOLLOWER MARKETS

In this section we analyze the Twitter follower markets that we discovered. In particular, we study the size of these markets and the price distribution for their offered services.

4.1 Market Prices

All the markets that we found offer Twitter followers for sale. Some of them offer additional services related to Twitter, such as having a tweet chosen by the customer retweeted a number of times. In addition, some of the markets do not only provide Twitter-related services but also target other social networks such as Facebook and Youtube.

The price for buying Twitter followers varies depending on the market, and ranges from $40 to $216 for 10,000 followers. Some markets go further, offering followers of a certain guaranteed "quality" (for example, each of the purchased followers will have 100 or more followers of their own). Intuitively, having "high quality" followers makes it possible to reach a broader audience, in case some of them retweet messages posted by the customer. The price for promotional tweets and retweets on the markets that we identified varies between $79 and $550 for 1,000 tweets (or retweets). In general, pyramid markets charge higher than non-pyramid markets as they can deliver real, compromised users as followers.

4.2 Market Sizes

We want to understand how prominent the Twitter follower markets are within the Twitter ecosystem, and in particular how many victims they are able to recruit. To answer this question, we searched for tweets advertising these markets in the Twitter stream. In particular, we had access to a random 10% sample of all public tweets for the period between January 16, 2013 and May 7, 2013. This accounted for 3.3 billion total tweets. To detect tweets advertising the markets in this stream, we proceeded as follows:

1. We group victim accounts in $\mathbf{A}_v$ based on the market that they belong to. For each market, we extract the set of victims of that particular market, which we call $\mathbf{A}_m$.
2. We group the tweets sent by the accounts in $\mathbf{A}_m$, based on text similarity. In particular, we group together those tweets that share four or more identical words.
3. For each group of tweets of size greater than one, we extract the URLs and hashtags contained in those tweets. We add these keywords to the set $\mathbf{K}_m$ of keywords indicative of the market m.
4. For each keyword in $\mathbf{K}_m$, we search the Twitter stream for tweets containing that word. We consider every hit as a tweet advertising the market m, and the account that posted the tweet as a victim of m.

The results are shown in Table 4. Although these results were derived from only 10% of the total tweets, the sheer volume of the victims is significant, confirming the impact of the follower markets on Twitter. Moreover, Twitter has a particularly hard time in dealing with this problem: since victims are usually legitimate accounts that got compromised, Twitter cannot just suspend them, but has to try alternative mitigation techniques, such as blocking the OAuth applications that are used by these markets.

Across the five identified markets, *Bigfolo* appears to be larger than the others, followed by *BigFollow* and *NewFollow*. *JustFollowers* appears to be considerably smaller. Our analysis on *InterTwitter* terminated at step 2 where we did not find similar tweets across victims advertising the market. Since *InterTwitter* does not offer a free subscription, it is likely that the victims are fake accounts with no real followers. Thus there is no point for victims to post advertisement tweets to attract more victims.

We then looked at the number of victims who advertise more than one market. These are users who, to get more "free" followers, handed out their credentials to multiple markets. In total, we found that only 22,702 out of 740K (or 3%) victims advertise more than one market.

We suspect that the size difference across these markets is the key reason of why the number of identified customers varies significantly across the markets. In particular, the fact that *Bigfolo* has so many more victims might mean that each victim follows a smaller number of customers, and therefore makes it harder for us to identify these customers. In Section 5.4, we will discuss more on how each follower merchant "distributes" their victims to service their customers.

5. ANALYZING CUSTOMERS AND VICTIMS

In this section, we analyze the characteristics of the key players of Twitter follower markets: market customers and victims, and we compare them to the characteristics of regular Twitter accounts. The goal is to identify behavioral features that can effectively distinguish between the follower purchasing phenomenon and an organic follower growth, typical of legitimate users.

5.1 Customer Account Characteristics

We first characterize the differences between market customers and legitimate accounts. We start by analyzing static characteristics of Twitter accounts, such as the current number of followers or friends that an account has. Figure 3 shows the cumulative distribution function (CDF) of the number of followers of market customers in $\mathbf{A}_c$ and 2 million randomly-sampled legitimate users $\mathbf{A}_{lr}$. Compared with regular Twitter users, market customers typically have more followers. The figure also indicates that customers typically have more than 100 followers.

Another interesting aspect is that, as Figure 4 shows, legitimate Twitter accounts typically have a more balanced follower-to-friend

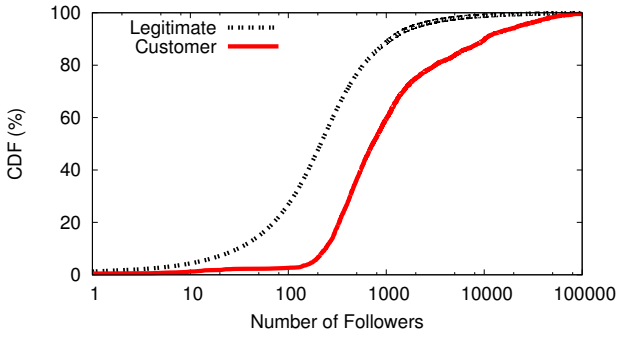


Figure 3: Cumulative distribution function of the number of followers of market customers, compared to legitimate users. Customers typically have one hundred followers or more.

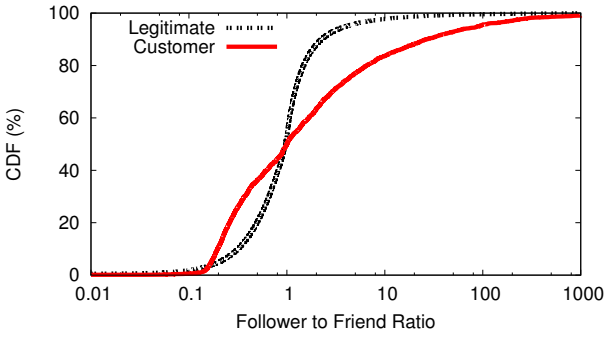


Figure 4: Cumulative distribution function of the follower to friend ratio of market customers, compared to legitimate users. 20% of market customers have at least ten times more followers than friends.

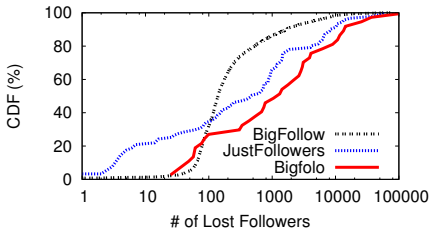


Figure 5: Cumulative distribution function of the number of followers that ever unfollowed the customer.

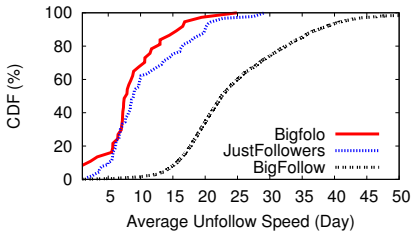


Figure 6: Cumulative distribution function of the average time it takes before followers unfollow the customer.

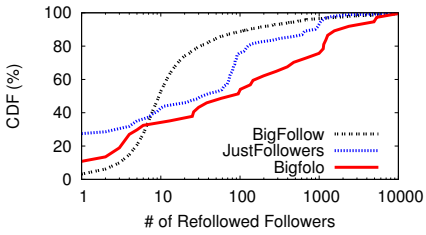


Figure 7: Cumulative distribution function of the number of followers that ever unfollow-then-refollow the customer.

ratio (*i.e.*, the number of followers an account has, divided by its number of friends) compared to accounts that purchased followers. In particular, 20% of the customers in $\mathbf{A}_c$ have at least ten times more followers than friends, while 98% of legitimate users in $\mathbf{A}_{lr}$ cannot achieve this. In addition, about 50% of the customers have more friends than followers. It turns out that despite buying followers, these customers also use other techniques to increase their number of followers, such as massively following random users with the expectation that some of them will follow back.

To analyze the people who followed the customers in $\mathbf{A}_c$, we crawled the follower list of the accounts in $\mathbf{A}_c$ on an hourly basis, from January 23, 2013 to May 7, 2013. We call the sequence of the followers for an account, crawled over multiple hours, *follower dynamics* for that account.

One of our hypothesis is that some market victims, who unwillingly followed a customer, would eventually unfollow the customer. So here, we try to verify whether customers in $\mathbf{A}_c$ lose followers over time. Figure 5 plots the distribution of the number of followers that ever unfollowed the accounts in $\mathbf{A}_c$. It shows that almost all the customers from the three markets have lost followers during our observation time. About 70% of customers have lost more than 100 followers. Figure 6 shows how long the "following" relationship will last before followers unfollow the accounts in $\mathbf{A}_c$. It shows that the *BigFollow* market provides followers with better loyalty, as 80% of the average "following" relationship can last at least 2 weeks. Victims of the other two markets usually unfollow the customer within 2 weeks.

We also observe that some followers who unfollowed customers would follow them again. Figure 7 shows the number of followers that ever unfollow-then-refollow the customer. This might indicate that some customers have bought the service more than once, and that the markets use the same set of victims for the same customer.

5.2 Follower Dynamics

We then wanted to understand whether there are differences in the way in which legitimate users and market customers acquire their followers over time. The datasets used to analyze follower dynamics are $\mathbf{A}_l$ and $\mathbf{A}_c$. Similarly to what we did for $\mathbf{A}_c$, the follower dynamics for the accounts in $\mathbf{A}_l$ were also collected by crawling their followers on an hourly basis, from January 23, 2013 to May 7, 2013. To this end, we first need to define a model for the dynamics of Twitter followers.

We define the current fluctuation in followers for an account a as

$$\Delta_a[f_h] = f_h - f_{h-1},$$

where h is the current observation period (of an hour). In a nutshell, $\Delta_a[f_h]$ represents the number of followers that account a gained (or lost) during hour h. Given this basic definition, we define four characteristics in the follower dynamics of an account: *increases in followers*, *steady decreases in followers*, *intervals of stationary followers*, and *follower fluctuations*.

Increases in Followers. Twitter accounts tend to increase the number of their followers over time. Generally, users experience a regular increase in followers, and sudden spikes in the number of followers are rare for legitimate users. On the other hand, those accounts that purchase followers might have many of these followers added in a short period of time. Given an account a, we say that the account experienced a follower increase of height t if, during

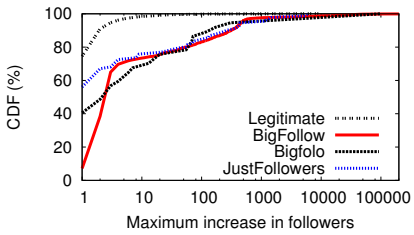


Figure 8: Cumulative distribution function of the sudden increase of followers experienced by customers, compared to legitimate profiles.

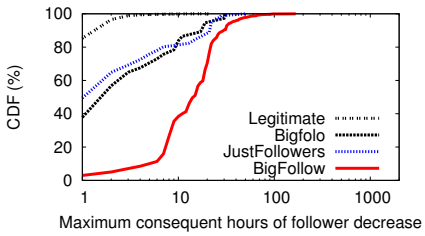


Figure 9: Cumulative distribution function of the maximum period of follower decrease experienced by customers, compared to legitimate profiles.

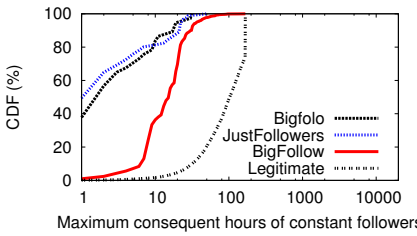
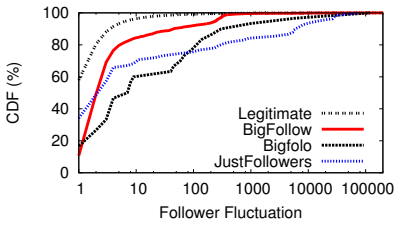
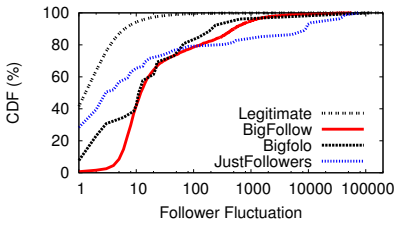


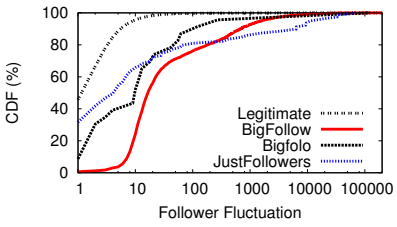
Figure 10: Cumulative distribution function of the maximum period of constant followers experienced by customers, compared to legitimate profiles.



(a) 1-hour Interval



(b) 12-hour Interval



(c) 24-hour Interval

Figure 11: Cumulative distribution function of the follower fluctuation with different interval lengths for market customers and legitimate users. Increasing the observation period, market customers show higher fluctuations than legitimate users.

the hour of observation, $\Delta_a[f_h]$ for the account is greater or equal than t.

To investigate these differences, we analyzed the dynamic characteristics of the users in $\mathbf{A}_l$ and $\mathbf{A}_c$. What we observed is that it is more common for market customers to experience a large increase in followers over a one-hour period than it is for legitimate users. Figure 8 shows this phenomenon. In particular, it is quite common for market customers (20%) to experience increases of 50 or more followers during a period of an hour, while only a small fraction (0.4%) of legitimate accounts experience the same.

Steady Decreases in Followers. Legitimate users get followed because they share interesting content, and engage their audience. Market customers, however, are usually trying to promote themselves or a brand, and the content that they share is not considered useful by many users. For this reason, many of the followers that a customer bought are likely to unfollow him after a certain time. We already verified this hypothesis in Section 5.1. Given an account a, we say that a experiences a steady decrease in followers of length d hours if, for a number d of consecutive hours, $\Delta_a[f_h]$ has been negative (*i.e.*, the account lost followers).

Figure 9 shows the CDF of the longest sequence of consecutive hours in which customers and legitimate users lost followers, over an observation period of one week. As it can be seen, the accounts in $\mathbf{A}_c$ show long periods in which they lose followers. In particular, 60% of the customers of *BigFollow* experienced periods of 10 or more consecutive hours in which their followers steadily decreased.

Intervals of Stationary Followers. Although Twitter accounts experience variations in their number of followers, we expect this number of followers to change slowly over time. In fact, most Twitter accounts will keep the same number of followers over periods of multiple hours, while market customers tend to constantly experience increases in followers or decreases in the number of their followers. For this reason, we expect legitimate Twitter accounts to show intervals of multiple hours in which their followers do not change more often than market customers. Given an account a, we say that a experienced an interval of constant followers of length d hours if, for a number d of consecutive hours, $\Delta_a[f_h]$ has been equal to zero.

Figure 10 shows the CDF of the longest sequence of consecutive hours in which market customers and legitimate users kept their followers unchanged, over an observation period of one week. As it can be seen, 30% of legitimate users did not experience changes in their followers at all. On the other hand, market customers rarely have their followers remain constant for longer than ten consecutive hours.

Follower Fluctuations. An important aspect in follower dynamics is the fluctuation of followers. Twitter accounts gain and lose followers, depending on how interesting the content that they share is, as well as other factors. In general, we expect the follower fluctuations of legitimate Twitter accounts to be rather small. On the other hand, accounts that bought followers are likely to show fluctuations that are more pronounced.

To investigate this phenomenon, we define the change in the number of followers over a period of n hours as

$$\Delta_a[f_{nh}] = f_h - f_{h-n}.$$

We then look for consecutive intervals of length n hours in which this variation had the same sign (positive or negative). For each interval in which the follower variation had the same sign, we calculate the total number of followers that the account gained (or lost) during that interval. By doing this, two consecutive intervals will have opposite signs, indicating, for example, an interval in which the followers of an account increased steadily, followed by an interval in which its followers decreased. Given two consecutive inter-

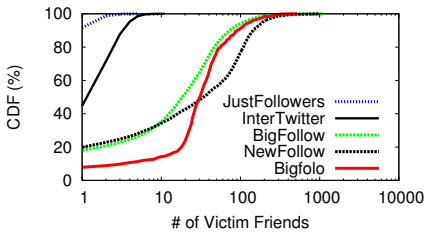


Figure 12: Cumulative distribution function of the number of other victims that market victims followed.

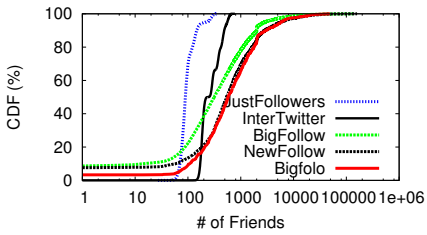


Figure 13: Cumulative distribution function of the total number of unique friends of victim accounts.

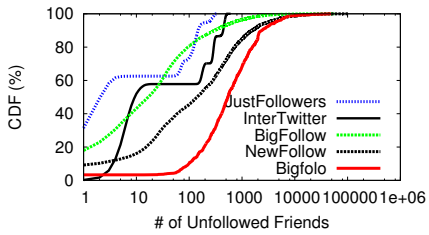


Figure 14: Cumulative distribution function of the number of friends that a victims unfollowed.

vals, we define the follower fluctuation Fl_n for periods of length n between those two intervals as the absolute value of the difference between the change of followers in the two intervals.

Intuitively, we expect accounts that have bought followers to show higher fluctuations than legitimate accounts. To verify this assumption, we calculated the fluctuation of followers for the accounts in $\mathbf{A}_l$ and $\mathbf{A}_c$ over a period length n of one hour, twelve hours, and 24 hours. For each account, we calculated the maximum value of Fl_n observed during the measurement period. Figure 11 shows the CDFs for these fluctuations. As it can be seen, legitimate accounts experience relatively low fluctuations, while customers tend to experience larger ones. Also, using longer time intervals (12 and 24 hours) leads to better distinction between the two groups of users. This indicates that both legitimate users and customers would experience short-terms fluctuations, while in the long term the fluctuations experienced by customers are more evident.

5.3 Victim Account Characteristics

In this section, we analyze the characteristics of victims in $\mathbf{A}_v$. We want to understand whether victims tend to follow other victims of the same market. Also, we want to verify whether victims would consistently follow and then unfollow other users.

In order to check the following behavior of users in $\mathbf{A}_v$, we crawled their friend list (*i.e.*, users they follow) every two hours from April 3, 2013 to April 30, 2013. During the data collection period, we observed that some of the victims got banned by Twitter. The victims of *InterTwitter* were suspended more frequently (31%) than the victims from other markets (*BigFollow* 8%, *Bigfolo* 11%, *JustFollowers* 17%, and *NewFollow* 23%). Note that *InterTwitter* is the only market among these five that does not offer a free-subscription service. It is likely that *InterTwitter* recruits victims by massively creating fake accounts, and existing countermeasures already detect such accounts. On the other hand, the remaining markets sell real accounts as followers. For Twitter, dealing with these accounts is more challenging, because blocking them might generate complaints by the accounts' owners.

To study the characteristics of victim accounts, we first want to see whether victims would follow each other. Figure 12 shows the distribution of the number of other victims that a victim followed. As shown, most victims of the markets *Bigfolo*, *BigFollow* and *NewFollow* would follow other victims. This way, victims can increase each other's follower count. However, the vast majority (more than 95% of them) of victims of *JustFollowers* followed at most one other victim from the same market. Victims of *InterTwitter* also show a low interconnection between victims: more than 90% of the victims have followed less than 3 other victims. Since *InterTwitter* does not offer free-subscriptions, victims do not have to follow each other. For *JustFollowers*, one possible explanation is that this market is considerably smaller compared to the others.

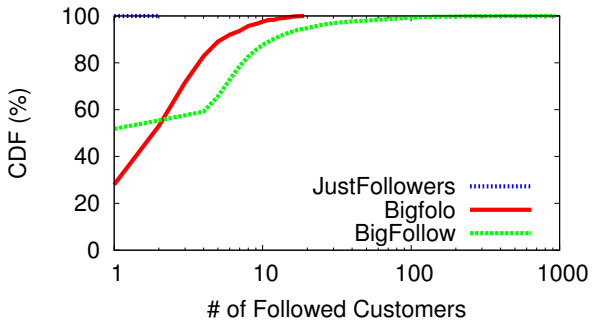


Figure 15: Number of customers that victims followed. Some markets have their victims follow many customers, while others have each victim follow a small number of customers.

This is confirmed by the number of victims for this market that we identified in the wild (see Table 4).

Figure 13 shows the total number of friends of the victim accounts in $\mathbf{A}_v$. This number is a cumulative count of all friends that we observed during the data collection period. We can see that victims of *InterTwitter* and *JustFollowers* strictly limit their number of friends below 500, while victims of the other three markets would follow more people. Normally, the friend count of a Twitter account cannot exceed 2,000, which is the hard limit set by Twitter. A small portion (about 10%) of victims of the markets *Bigfolo*, *BigFollow* and *NewFollow* have followed more than 2,000 people during one month. The possible explanation is that victims would unfollow some of their friends in order to follow new ones. This can be confirmed in Figure 14, which shows the number of friends that the victims in $\mathbf{A}_v$ unfollowed during the data collection period. As we can see, market victims typically unfollow many of their friends. For example, 90% of the victims of *Bigfolo* have unfollowed more than 100 people during the period that we observed, while 30% of victims have unfollowed more than 1,000 people. Previous research showed that it is not uncommon for legitimate Twitter users to unfollow other profiles, but at rates a lot lower than this [22].

5.4 Different Strategies for Operating Markets

As we have observed, the Twitter follower markets that we have analyzed share common traits, but also feature distinguished characteristics. In this section, we study the different strategies that follower merchants use to operate their markets, and analyze advantages and disadvantages of these strategies.

The most evident difference is that the followers offered by *InterTwitter* seem to be fake accounts, unlike those of all other mar-

kets, which are real accounts. The first consequence of this choice is that such followers will not provide an increase in popularity to their customers, but just an increase in their follower count. Fake accounts will not purchase the products advertised by the customer, for instance. In addition, as we have shown, it is easier for Twitter to deal with fake accounts, because the social network can just suspend them, without anybody complaining. The follower merchants can create new fake accounts, but this is not effortless.

The main difference between the other three markets (*BigFollow*, *Bigfolo*, and *JustFollowers*) is in the way they manage their victims. First of all, as we showed in Section 4.2, some markets can count on more victims than others. Intuitively, the victims of the larger markets will have less connections to the customers of the market, meaning that each victim will follow a smaller number of customers. Figure 15 shows the CDF of how many accounts in $\mathbf{A}_c$ are followed by the victims in $\mathbf{A}_v$. As it can be seen, the victims of *JustFollowers* follow a very small number of customers, while the ones of *Bigfolo* and *BigFollow* follow a larger number of them. Another way in which follower merchants can control the way in which they deliver followers is how fast they add them: as we have shown in Figure 8, *BigFollow* provides customers in a more abrupt way than *Bigfolo* and *JustFollowers*. Another interesting aspect is whether follower merchants provide the same followers twice, in case the customer purchased their service more than once. As we show in Figure 7, *JustFollowers* and *Bigfolo* tend to give the same followers to their users more than once more often than *BigFollow*. Given the number of tweets advertising these two markets that we observed in the wild, and that are summarized in Table 4, the reason for this is probably that *Bigfollow* can count on a larger number of victims, while *JustFollowers* is a lot smaller.

6. DETECTING MARKET CUSTOMERS

Previously we analyzed the characteristics of market victims and customers. Some of these characteristics can be leveraged to detect accounts that purchased followers on Twitter (i.e., market customers). Detecting customers is very important to fight against the phenomenon of Twitter follower markets: while crooks can compromise more accounts and sell them as followers, it is much harder for them to attract new customers, especially if Twitter starts to systematically suspend customer accounts. In principle, Twitter could start suspending customer accounts, because they are violating the terms of service that specifically prohibit to take part in Twitter follower market activity [3]. However, to date we are not aware of any countermeasures taken by Twitter against market customers.

We propose a method that leverages the dynamics of the followers of an account, and raises an anomaly if they show the patterns that we described in Section 5.2. We also explore different filters that can help in discarding many accounts that are unlikely to be customers, and a classifier that can tell if an account is likely a market customer just by looking at its static characteristics. Eventually, we discuss how these techniques can be used in combination, and leverage them to detect customers in the wild.

6.1 Follower Dynamics Detection

As explained in Section 5.2, Twitter follower market customers tend to experience sudden increases in their followers, and steady decreases of the number of their followers for long periods of time. We are interested in leveraging these observations to detect accounts whose owners bought followers on Twitter (i.e., customers). To this end, we observe the follower dynamics of Twitter accounts for a number d of observation periods (hours), and we leverage the follower dynamics that are typical of accounts that bought followers to detect customers in the wild. Based on our observations, we developed three types of features to perform this task: *increase* features, *decrease* features, and *stationary followers* features. We describe them in detail in the following.

Increase Features. These features count the number of times during the observation period in which an account experienced an increase in followers higher or equal than t during one hour. We define 1,000 features of this type, with t ranging from 1 to 1,000.

Decrease Features. These features count the number of periods of length l hours in which the followers of a user steadily decreased. We define d features of this type, where the length of the steady decrease ranges from 1 hour to d hours.

Stationary Features. These features are similar to the steady decrease features, but look for consecutive hour intervals in which the user's number of followers did not change. Again, we define d features, from 1 hour to d hours.

In Section 5.2 we defined another characteristic of the follower dynamics of an account, the *follower fluctuation*. However, this characteristic embodies the same information provided by the *increase* and *decrease* features. In particular, if we look for accounts that ever experienced a sudden increase of followers of height greater or equal than 15, or that experienced a steady decrease in followers that lasted for at least ten hours, we already cover 97.6% of the accounts in $\mathbf{A}_c$. On the other hand, only 1% of the accounts in $\mathbf{A}_l$ show the same characteristics. Since the follower fluctuation is more resource-intensive to compute than the increase and decrease features, we decided not to use this type of features for our classification.

When analyzing follower dynamics, we have to observe Twitter accounts for a certain period of time. One of the challenges is to determine the optimum length of the observation period. If we picked an observation period that is too short, our system would probably not be able to observe the dynamics typical of market customers during that period. Also, the dynamics that are typical of a purchase of followers do not last forever, but instead tend to stabilize after a certain point. For this reason, to successfully detect a customer, we need the purchase to happen during our observation period, or shortly before it. We experimented with different observation periods, and found that having an observation period of one week (d =168 hours) is a good choice. In the next section, we perform an analysis of the classifier built from these features.

6.1.1 Dynamics-based Classifier Evaluation

We built a classifier based on the follower dynamics features, and analyzed its efficiency. To this end, leveraged Support Vector Machines trained with Sequential Minimal Optimization (SMO) for classification [27] as provided by the WEKA machine learning toolkit [15].

For our training set, we leveraged the set $\mathbf{A}_c$ as examples of accounts that bought followers, and a random set of 10,000 accounts from $\mathbf{A}_l$ as examples of legitimate accounts. We call this set $\mathbf{A}_{lt}$. We then extracted the follower dynamics for the accounts in our training set, for the period between April 16, 2013 and April 23, 2013. A ten-fold cross validation on this dataset shows that the classifier works well. We obtained a true positive rate of 98.4% among the market customers, and a false positive rate close to zero: only two of the profiles in $\mathbf{A}_{lt}$ were flagged as customers. The reason for this is that both accounts experienced several high increases in followers during the observation period. By manually analyzing these accounts, one was a foreign news site, and is probably a false positive, while the other one is belongs to a local DJ. We cannot tell if this particular account is a false positive or not.

The features that are most representative of a market customer are high increases in followers. However, it is very rare for a customer to experience an increase of more than 200 followers during an hour. Long intervals of steady decreases in followers (longer than ten consecutive hours) are also indicative of customers. Surprisingly however, shorter intervals (one or two hours) of steady decreases in followers are more common in legitimate accounts. The stationary features have less influence on the model.

Although this classifier is very strong, a market might try to avoid detection, by adding followers slowly. We argue that while customers of such a market might not look suspicious with regard to the features that deal with increases in followers, they will still show anomalous steady decreases in followers, because this is not an element that the follower merchants have control over.

6.2 Static Detection of Customers

We showed that it is possible to reliably detect accounts that bought followers on Twitter by looking at their follower dynamics. However, this method requires to observe an account for a long period of time. In our case in particular, it is unfeasible to monitor the whole population of Twitter, because Twitter limits us in the number of API calls that we can perform per hour. To be able to run our customer detection system at Twitter scale, we would need a lightweight system that discards as many benign users as possible before we start tracking the possible customers, to analyze their dynamics. To this end, we analyzed three possible strategies: using a *follower filter*, a *static filter*, or a *static classifier*. In the following, we analyze the three methods in detail.

Follower Filter. As Figure 3 illustrates, almost all market customers have more than 100 followers. Conversely, 26% of legitimate Twitter users are below that threshold. This shows that a quick way of getting rid of profiles that are unlikely to be customers is considering only accounts that have more than a hundred followers. In practice, this is what we already did by using the dataset $\mathbf{A}_l$ as our test dataset.

Static Filter. A more advanced way of filtering out accounts that are unlikely to be market customers is looking at static characteristics of a profile, such as the number of followers and the number of friends. The idea is to leverage the observations that we discussed in Section 5.1 to identify possible customer candidates. In the following, we describe the features used by this static filter in detail.

- **Number of Friends.** This is the number of friends of the Twitter account.
- **Number of Followers.** As we previously explained, market customers have, on average, more followers than normal Twitter users.
- **More Followers than Friends.** As we show in Figure 4, it is slightly more likely for market customers to have more followers than friends than it is for legitimate users. This feature is Boolean, and it is set to one if the number of followers is higher than the number of friends.
- **Followers to Friends Ratio.** As we show in Figure 4, customer accounts tend to have a higher follower-to-friend ratio than legitimate users.

This filter has the advantage of being fast, and being able to quickly discard accounts that are most probably not market customers. However, since it only takes into account the number of friends and followers of an account, it might generate many false positives. Since we use this filter to discard accounts that are not likely to be market customers, before starting to monitor account dynamics, this is not a big problem.

Dynamic Classifier	TP rate	FP rate
SVM (SMO)	98.4%	0.02%
Static Filter	TP rate	FP rate
Random Forest (cost-sensitive)	93.7%	63%
Static Classifier	TP rate	FP rate
Decision Tree	90%	3.7%
Random Forest	91%	3.3%

Table 5: Performance of the different classifiers. The dynamic classifier has both high recall and low false positives. The static methods are useful as prefilters to discard as many accounts that are unlikely to be market customers as possible.

Static Classifier. To be more precise in assessing whether an account is likely a market customer or not, we can add two additional features to the static filter.

- **Influence.** For a legitimate account, having a high number of followers means that people find this account interesting, and re-share the content that the account posts often. This indicates that such an account has high influence. Accounts that bought their followers, typically do not have very engaged audiences. We used the *Klout* service to measure an account influence [19]. This service analyzes the activity of Twitter accounts, and returns an influence score that is function of how many followers, mentions, and retweets an account has. This score performs well in our case. However, any service that provided a similar information would work for our purposes.
- **Number of Victim Followers.** Since Twitter follower markets use compromised accounts as followers for their customers, we expect market customers to be followed by many victims. To calculate this feature, for each follower b, of an account a, we look for tweets advertising an account market, similar to what we did in Section 3.2. We set this feature to the number of followers that follow a and have advertised a Twitter follower market.

The static classifier is more robust than the static filter, because it takes into account information about the activity of the account, as well as its social network. However, as we will see, it is not as lightweight.

In the following, we analyze the performance of the static filter and the static classifier, and describe their advantages and disadvantages.

6.2.1 Evaluation of the Static Methods

We performed an analysis of the two methods listed in the previous section: the static filter and the static classifier. Similarly to what we did for the follower dynamics classifier, we used the accounts in $\mathbf{A}_c$ as examples of market customers, and the accounts in $\mathbf{A}_{lt}$ as examples of legitimate users.

Our goal with the static filter is to discard as many benign accounts as possible, and analyze the remaining ones with the dynamic classifier. To perform this task, we require recall to be high, but allow false positives. First, we performed a 10-fold cross validation on the training dataset for the static filter. To this end, we used random forests, and we penalized a false negative 100 times

more than a false positive. The 10-fold cross validation returned a true positive rate of 93.7%, and a false positive rate of 63%. As we can see, the static filter is able to detect most customer accounts, and is able to cut the number of candidates for the dynamic classifier by almost a half. Therefore, it is a useful pre-filter.

We then wanted to understand how much the features added to the static classifier improve over the simple filter, and whether performing this type of classification could be useful. To this end, we first performed a 10-fold cross validation, using decision trees. The 10-fold cross validation returned a true positive rate of 90%, and a false positive rate of 3.7%. The most important features in the decision tree were, in order, the number of followers, whether the account has more friends than followers, and the influence score. A random forest algorithm gave slightly better results [24], with a true positive rate of 91%, and a false positive rate of 3.3%. The results that we obtained for different approaches are summarized in Table 5.

The 10-fold cross validation on the static classifier shows that the features that we developed are effective in discriminating between users who acquired followers in a legitimate fashion and those who purchased followers. In principle, we could use this approach to detect market customers in the wild. Unfortunately, from our perspective, Twitter limits the number of API calls that we can make in an hour. Some of the features that we described, in particular the "Number of Victim Followers" and the "Influence" ones, require many API calls to compute, since they require us to download the timeline for every single follower that an account has, as well as the past tweets of the user. For these reasons, although we acknowledge that the static classifier might help in detecting customers in the wild, we did not use it to this end.

A market customer might evade detection by the static classifier by using other means of getting followers in conjunction with purchasing them. For example, a customer who constantly follows random users, and gets followed by a fraction of them, would have a more balanced number of friends and followers than average market customers, and might avoid detection. However, Twitter is already monitoring and blocking accounts that show this behavior [32].

6.3 Possible Uses of the Two Methods

We presented a method to monitor the follower dynamics of Twitter accounts, and determine if they increased their followers organically, or they purchased followers. This method works very well, but is not instantaneous, since it requires a long observation period before making a decision (in our setup, a week). Also, monitoring the follower dynamics of Twitter accounts is resource intensive. To mitigate these problems, we developed two methods to discard those accounts that are unlikely to be market customers (*follower filter* and *static filter*), and a method to tell if a profile is likely a market customer by only looking at it (*static classifier*). There are different ways in which these techniques can be combined effectively.

The most reasonable deployment would be applying the follower filter, followed by the static classifier first. Both methods are fast, and work well in discarding unlikely customers. Then, we can apply the dynamics classifier to the remaining set of accounts. We investigate this possible setup in the next section. Another possible deployment is to use the dynamic and the static classifier in parallel. As we have mentioned in the previous sections, both classifiers might be actively evaded by follower merchants and customers, and having two detection mechanisms in place might help in detecting these evasive accounts.

6.4 Detecting Customers in the Wild

We ran our detection algorithms on $\mathbf{A}_l$. In particular, we extracted the follower dynamics for the 2 million accounts for a period of two weeks, from April 19, 2013 to May 3, 2013.

First, we ran our dynamic classifier over the two-week data, on each week separately. Our system flagged 684 accounts as market customers. As we explained previously, this means that these accounts purchased followers during our observation period, or right before that, and that they showed follower dynamics that are typical of market customers. By manually looking at those accounts, they mostly belonged to wanna-be celebrities or small businesses, which is the type of accounts that we would expect to purchase Twitter followers to bootstrap their popularity. Although it is not possible to establish precise ground truth for our results, none of the detected customers looked like a false positive (for example, no account belonged to an established celebrity or a popular business).

We then looked at how applying the filter methods that we discussed in Section 6.3 would have affected our results. As we have shown in Figure 3, applying the follower filter would discard between 20 and 30% legitimate Twitter accounts, while keeping the vast majority of customers in the candidate set. For our purposes, we do not need to apply this filter, since the accounts in $\mathbf{A}_l$ have been selected to have more than 100 followers. We then applied the static filter to the accounts in $\mathbf{A}_l$. In total, the filter discarded 983,810 accounts, almost cutting the candidate set in half. After applying this filter, we ran the dynamics classifier on the remaining accounts. Our system flagged 631 accounts as market customers. This means that the static filter is fairly successful in keeping most customers in the candidate set – 92.3% of the total detected customers were still in it – and shows that having a static filter is a good option if the number of accounts that one can monitor is limited (like in our case).

6.5 Analysis of the Identified Customers

In this section, we analyze the characteristics of the 684 market customers that we identified. These customers are detected by the dynamic classifier, so not surprisingly, their dynamic characteristics are very similar to the ones of the customers in the training dataset $\mathbf{A}_c$. We found that their static characteristics show strong customer-like signals too: Figure 16 shows the distribution of the number of followers of the identified customers, compared to legitimate users. The identified customers typically have more than 1,000 followers, which is more than the number of followers for 80% of the regular Twitter user population. In addition, Figure 17 shows that the follower-to-friend ratio of identified customers is typically higher than one.

As we said, by manually looking at the identified customers we found that those accounts mostly belong to small businesses or wannabe celebrities, who try to boost their popularity. We then wanted to assess whether the purchase of followers actually helps in this process. To this end, we analyzed the influence score of these customers, according to Klout [19]. The CDF of the influence of the identified customers is shown in Figure 18, indicating that about half the customers have a Klout score lower than 45. However, the median Klout score of Twitter accounts is precisely 45 [1] (on a scale 1-100). This shows that, although purchasing followers can boost an account's social network, it does not really help in making the account popular. Since the followers did not willingly follow the profile, it is unlikely that they will engage the user, and share her content.

As a last element, we wanted to understand whether Twitter is detecting and blocking these customer accounts. After one week from detecting them, only two accounts had been suspended by

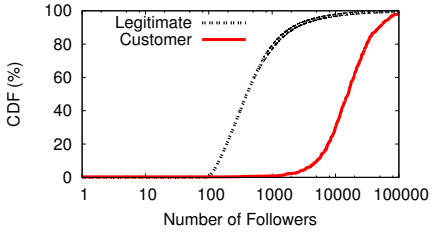


Figure 16: Followers of identified customers and legitimate users.

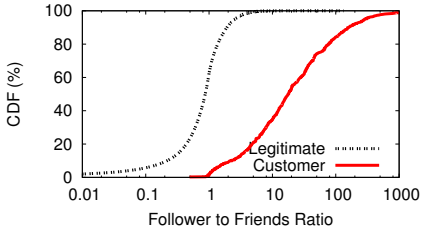


Figure 17: Follower-friend ratio of identified customers and legitimate users.

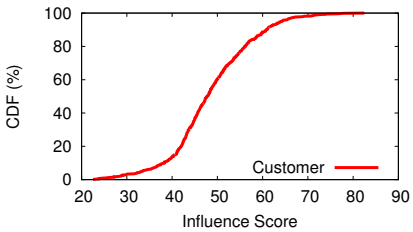


Figure 18: Influence (Klout) score of identified customers.

Twitter. This shows that it is hard for Twitter to detect which accounts purchased followers, and that the type of techniques proposed in this paper could actively help Twitter in fighting this phenomenon.

7. DISCUSSION

Follower markets enable dishonest users on social networks and microblogging sites to inflate their perceived credibility. Due to the lack of alternative means to gauge a user's influence, social network users often take the number of connections of an account as face-value for the account's influence. The techniques proposed in this work can help the social network operator to automatically detect the participants in these follower markets. More importantly, our system distinguishes between customers and victims of these schemes. This distinction allows the social network operator to focus mitigation approaches on market customers and potentially disrupt the economic foundations of follower market services. Furthermore, unsuspecting victims need not be punished and their participation in the social network can continue. As a consequence, the intuition that accounts with many followers are influential is restored.

Additionally, our techniques can also be leveraged by users that strive to keep their online social networking relations confined to trustworthy peers. To this end, a user can employ our techniques to identify market customers among the users in her social graph. Due to the high precision of our detection system, the user can remove the identified market customers from her social graph without fear that she would break connections to legitimate users in the social network.

Of course, once our detection mechanism impacts the bottom line of follower markets, follower merchants might try to adapt their operations to evade our techniques. However, our detection techniques leverage a key observation about the victims of these services – the fact that many victims unfollow the customers the merchants connect them to. Therefore, as long as follower markets exploit real victims for their operations we expect that these characteristics will persist. Follower merchants can still slow down the rate at which victims are added to customers to evade detection. While this would be a viable long term strategy for the follower merchants, it dramatically impacts one of the markets' key promises – to gain followers fast. These considerations do not apply to those follower markets that use fake accounts instead of real ones. However, this modus operandi has higher operational costs and overhead for the follower merchants, because they have to create and manage a substantial number of fake accounts. Such an operation by itself is non-trivial at best and would expose these fake accounts to existing detection techniques already employed by today's online social networks.

8. RELATED WORK

Over the last few years, many researchers focused their studies on Twitter, and online social networks [8, 20–22, 35, 36]. In particular, several works showed the security threats linked to the use of social networks [12–14]. Previous research showed that one of the main advantages of spreading malicious content through social networks is that users are more likely to click on content posted on these media than they are of doing so for more traditional media (*e.g.*, email) [6, 17].

Researchers developed several systems to detect malicious activity on social networks such as Twitter. Early systems focused on spammers and fake accounts (*Sybils*), and developed feature-based detection systems [5, 23, 29]. Others rely on the social network structure and detect clustered malicious accounts in the social graph [7, 37]. These techniques give a good first-level defense against social network threats.

More recently, miscreants started to send malicious content from legitimate accounts that had been compromised [12, 14]. Detecting compromised accounts is a very different problem compared to detecting fake ones, because such accounts can have a long history of legitimate activity, and typically do not show similar characteristics. Existing systems either look at the URLs that the malicious messages point to [30], at accounts that send the same malicious content in large-scale campaigns [12], or look for accounts that suddenly change their typical behavior [11].

Most existing detection techniques focus on the actual accounts sending out offending content. Once detected, social network administrators can suspend these malicious accounts. However, this modus operandi does not fundamentally remove the threats, as more malicious accounts could be created, or other legitimate profiles could get compromised. In this paper, we propose to detect and disrupt Twitter follower market operations by identifying and possibly suspending the accounts that fund these malicious enterprises – the markets' customers. By doing this, we hope to cause a major economical hit to the Twitter follower market schemes, up to a point in which they will not be profitable anymore.

Thomas et al. studied marketplaces where miscreants can buy fake accounts, and use them to spread malicious content [31]. Instead, we focus on markets that sell followers to their customers. The concept of Twitter follower markets was first introduced in our previous work [28]. In this paper, we studied the phenomenon more broadly, and focused on characterizing the behavioral patterns of key players of account markets. In particular, we leveraged the follow/unfollow dynamics to detect market customers, which the previous approach did not explore. Kwak et al. studied the unfollow dynamics on Twitter with a focus on normal users [21], while our study leveraged unfollow dynamics for anomaly detection from a security angle.

9. CONCLUSIONS

Our work studies Twitter follower markets, where followers are sold to customers for a fee. We use ground-truth to reveal interesting patterns in the dynamics of the users involved in these markets, and in particular, the follower populations of their customers. We use our insights to build a system for detecting these behaviors in the wild, and use real experiments to show that it is both scalable and robust.

10. ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research (ONR) under Grant N000140911042, the Army Research Office (ARO) under grant W911NF0910553, the National Science Foundation (NSF) under grant CNS-0845559 and grant CNS-0905537, and Secure Business Austria. Gang Wang, Haitao Zheng, and Ben Y. Zhao were partially supported by NSF grants IIS-0916307, CNS-1224100, IIS-1321083, and DARPA grant BAA-12-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

We would like to thank our shepherd Cecilia Mascolo for her support, and the anonymous reviewers for their insightful comments and observations.

11. REFERENCES

[1] Klout blog. http://corp.klout.com/blog/2011/11/do-you-have-google-klout/?goback=%2Egmp_4159546%2Egde_4159546_member_212786154.

[2] Oauth community site. http://oauth.net.

[3] Twitter terms of service. http://support.twitter.com/articles/18311-the-twitter-rules.

[4] Justin Bieber Twitter followers 50% fake says report, 2013.

[5] BENVENUTO, F., MAGNO, G., RODRIGUES, T., AND ALMEIDA, V. Detecting Spammers on Twitter. In *Conference on Email and Anti-Spam (CEAS)* (2010).

[6] BILGE, L., STRUFE, T., BALZAROTTI, D., AND KIRDA, E. All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In *World Wide Web Conference (WWW)* (2009).

[7] CAI, Z., AND JERMAINE, C. The latent community model for detecting sybils in social networks. In *Symposium on Network and Distributed System Security (NDSS)* (2012).

[8] CHA, M., HADDADI, H., BENVENUTO, F., AND GUMMADI, K. Measuring User Influence in Twitter: The Million Follower Fallacy. In *International AAAI Conference on Weblogs and Social Media* (2010).

[9] COLDEWEY, D. Romney twitter account gets upsurge in fake followers, but from where? *NBC News* (2012).

[10] CONSIDINE, A. Buying their way to twitter fame. *The New York Times* (2012).

[11] EGELE, M., STRINGHINI, G., KRUEGEL, C., AND VIGNA, G. COMPA: Detecting Compromised Accounts on Social Networks. In *Symposium on Network and Distributed System Security (NDSS)* (2013).

[12] GAO, H., HU, J., WILSON, C., LI, Z., CHEN, Y., AND ZHAO, B. Detecting and Characterizing Social Spam Campaigns. In *ACM SIGCOMM Internet Measurement Conference* (2010).

[13] GHOSH, S., VISWANATH, B., KOOTI, F., SHARMA, N. K., KORLAM, G., BENEVENUTO, F., GANGULY, N., AND GUMMADI, K. P. Understanding and combating link farming in the twitter social network. In *World Wide Web Conference (WWW)* (2012).

[14] GRIER, C., THOMAS, K., PAXSON, V., AND ZHANG, M. @spam: the underground on 140 characters or less. In *ACM Conference on Computer and Communications Security (CCS)* (2010).

[15] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. H. The weka data mining software: an update. *SIGKDD Explor. Newsl. 11*, 1 (Nov. 2009), 10–18.

[16] HUTTO, C. J., YARDI, S., AND GILBERT, E. A longitudinal study of follow predictors on twitter. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)* (2013).

[17] JAGATIC, T., JOHNSON, N., JAKOBSSON, M., AND JAGATIF, T. Social phishing. *CACM 50*, 10 (2007), 94–100.

[18] KING, R. How companies use Twitter to bolster their brands. In *BusinessWeek Online* (2008).

[19] klout. http://klout.com.

[20] KRISHNAMURTHY, B., GILL, P., , AND ARITT, M. A few chirps about twitter. In *USENIX Workshop on Online Social Networks* (2008).

[21] KWAK, H., CHUN, H., AND MOON, S. Fragile online relationship: a first look at unfollow dynamics in twitter. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)* (2011).

[22] KWAK, H., LEE, C., PARK, H., AND MOON, S. What is Twitter, a social network or a news media? In *World Wide Web Conference (WWW)* (2010).

[23] LEE, K., CAVERLEE, J., AND WEBB, S. Uncovering social spammers: social honeypots + machine learning. In *International ACM SIGIR Conference on Research and Development in Information Retrieval* (2010).

[24] LIAW, A., AND WIENER, M. Classification and regression by randomforest. *R news* (2002).

[25] MOTOYAMA, M., MCCOY, D., LEVCHENKO, K., SAVAGE, S., AND VOELKER, G. An analysis of underground forums. In *ACM SIGCOMM Internet Measurement Conference* (2011).

[26] PERLROTH, N. Fake twitter followers become multimillion-dollar business. *The New York Times* (2013).

[27] PLATT, J. C. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning* (1998).

[28] STRINGHINI, G., EGELE, M., KRUEGEL, C., AND VIGNA, G. Poultry Markets: On the Underground Economy of Twitter Followers. In *SIGCOMM Workshop on Online Social Networks* (2012).

[29] STRINGHINI, G., KRUEGEL, C., AND VIGNA, G. Detecting Spammers on Social Networks. In *Annual Computer Security Applications Conference (ACSAC)* (2010).

[30] THOMAS, K., GRIER, C., MA, J., PAXSON, V., AND SONG, D. Design and Evaluation of a Real-Time URL Spam Filtering Service. In *IEEE Symposium on Security and Privacy* (2011).

[31] THOMAS, K. AND MCCOY, D. AND GRIER, C. AND KOLCZ, A. AND PAXSON, V. Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse. In *USENIX Security Symposium* (2013).

[32] TWITTER. The twitter rules. http://support.twitter.com/entries/18311-the-twitter-rules, 2010.

[33] WANG, G., KONOLIGE, T., WILSON, C., WANG, X., ZHENG, H., AND ZHAO, B. Y. You are how you click: Clickstream analysis for sybil detection. In *USENIX Security Symposium* (Washington, DC, 2013).

[34] WANG, G., MOHANLAL, M., WILSON, C., WANG, X., METZGER, M., ZHENG, H., AND ZHAO, B. Y. Social Turing tests: Crowdsourcing sybil detection. In *Symposium on Network and Distributed System Security (NDSS)* (2013).

[35] WILSON, C., BOE, B., SALA, A., PUTTASWAMY, K., AND ZHAO, B. User Interactions in Social Networks and Their Implications. In *ACM European conference on Computer systems (EuroSys)* (2009).

[36] YANG, C., HARKREADER, R., AND GU, G. Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers. In *Symposium on Recent Advances in Intrusion Detection (RAID)* (2011).

[37] YU, H., KAMINSKY, M., GIBBONS, P., AND FLAXMAN, A. Sybilguard: defending against sybil attacks via social networks. *ACM SIGCOMM Computer Communication Review* (2006).