Detecting Spammers on Social Networks

Gianluca Stringhini
University of California, Santa Barbara
gianluca@cs.ucsb.edu

Christopher Kruegel
University of California, Santa Barbara
chris@cs.ucsb.edu

Giovanni Vigna
University of California, Santa Barbara
vigna@cs.ucsb.edu

ABSTRACT

Social networking has become a popular way for users to meet and interact online. Users spend a significant amount of time on popular social network platforms (such as Facebook, MySpace, or Twitter), storing and sharing a wealth of personal information. This information, as well as the possibility of contacting thousands of users, also attracts the interest of cybercriminals. For example, cybercriminals might exploit the implicit trust relationships between users in order to lure victims to malicious websites. As another example, cybercriminals might find personal information valuable for identity theft or to drive targeted spam campaigns.

In this paper, we analyze to which extent spam has entered social networks. More precisely, we analyze how spammers who target social networking sites operate. To collect the data about spamming activity, we created a large and diverse set of "honey-profiles" on three large social networking sites, and logged the kind of contacts and messages that they received. We then analyzed the collected data and identified anomalous behavior of users who contacted our profiles. Based on the analysis of this behavior, we developed techniques to detect spammers in social networks, and we aggregated their messages in large spam campaigns. Our results show that it is possible to automatically identify the accounts used by spammers, and our analysis was used for take-down efforts in a real-world social network. More precisely, during this study, we collaborated with Twitter and correctly detected and deleted 15,857 spam profiles.

1. INTRODUCTION

Over the last few years, social networking sites have become one of the main ways for users to keep track and communicate with their friends online. Sites such as Facebook, MySpace, and Twitter are consistently among the top 20 most-viewed web sites of the Internet. Moreover, statistics show that, on average, users spend more time on popular social networking sites than on any other site [1]. Most social networks provide mobile platforms that allow users to access their services from mobile phones, making the access to these sites ubiquitous.

The tremendous increase in popularity of social networking sites allows them to collect a huge amount of personal information about the users, their friends, and their habits. Unfortunately, this wealth of information, as well as the ease with which one can reach many users, also attracted the interest of malicious parties. In particular, spammers are always looking for ways to reach new victims with their unsolicited messages. This is shown by a market survey about the user perception of spam over social networks, which shows that, in 2008, 83% of the users of social networks have received at least one unwanted friend request or message [16].

From a security point of view, social networks have unique characteristics. First, information access and interaction is based on trust. Users typically share a substantial amount of personal information with their friends. This information may be public or not. If it is not public, access to it is regulated by a network of trust. In this case, a user allows only her friends to view the information regarding herself. Unfortunately, social networking sites do not provide strong authentication mechanisms, and it is easy to impersonate a user and sneak into a person's network of trust [15]. Moreover, it often happens that users, to gain popularity, accept any friendship request they receive, exposing their personal information to unknown people. In other cases, such as MySpace, the information displayed on a user's page is public by design. Therefore, anyone can access it, friend or not. Networks of trust are important from a security point of view, because they are often the only mechanism that protects users from being contacted by unwanted entities.

Another important characteristic of social networks is the different levels of user awareness with respect to threats. While most users have become aware of the common threats that affect the Internet, such as e-mail spam and phishing, they usually do not show an adequate understanding of the threats hidden in social networks. For example, a previous study showed that 45% of users on a social networking site readily click on links posted by their "friend" accounts, even if they do not know that person in real life [10]. This behavior might be abused by spammers who want to advertise web sites, and might be particularly harmful to users if spam messages contain links to malicious pages.

Even though social networks have raised the attention of researchers, the problem of spam is still not well understood. This paper presents the results of a year-long study of spam activity in social networks. The main contributions of this paper are the following:

ACSAC '10 Dec. 6-10, 2010, Austin, Texas USA
Copyright 2010 ACM 978-1-4503-0133-6/10/12 ...$10.00.

- We created a set of honeynet accounts (honey-profiles) on three major social networks, and we logged all the activity (malicious or not) these accounts were able to observe over a one-year period for Facebook and an eleven-month period for Twitter and MySpace.

- We investigate how spammers are using social networks, and we examine the effectiveness of the countermeasures that are taken by the major social network portals to prevent spamming on their platforms.

- We identify characteristics that allow us to detect spammers in a social network.

- We built a tool to detect spammers, and used it on a Twitter and Facebook dataset. We obtained some promising results. In particular, we correctly detected 15,857 on Twitter, and after our submission to the Twitter spam team, these accounts were suspended.

2. BACKGROUND AND RELATED WORK

Social networks offer a way for users to keep track of their friends and communicate with them. This network of trust typically regulates which personal information is visible to whom. In our work, we looked at the different ways in which social networks manage the network of trust and the visibility of information between users. This is important because the nature of the network of trust provides spammers with different options for sending spam messages, learning information about their victims, or befriending someone (to appear trustworthy and make it more difficult to be detected as a spammer).

2.1 The Facebook Social Network

Facebook is currently the largest social network on the Internet. On their website, the Facebook administrators claim to have more than 400 million active users all over the world, with over 2 billion media items (videos and pictures) shared every week [3].

Usually, user profiles are not public, and the right to view a user's page is granted only after having established a relationship of trust (paraphrasing the Facebook terminology, *becoming friends*) with the user. When a user A wants to become friend with another user B, the platform first sends a request to B, who has to acknowledge that she knows A. When B confirms the request, a friendship connection with A is established. However, the users' perception of Facebook friendship is different from their perception of a relationship in real life. Most of the time, Facebook users accept friendship requests from persons they barely know, while in real life, the person asking to be friend would undergo more scrutiny.

In the past, most Facebook users were grouped in *networks*, where people coming from a certain country, town, or school could find their neighbors or peers. The default privacy setting for Facebook was to allow all people in the same network to view each other's profiles. Thus, a malicious user could join a large network to crawl data from the users on that network. This data allows an adversary to carry out targeted attacks. For example, a spammer could run a campaign that targets only those users whose profiles have certain characteristics (e.g., gender, age, interests) and who, therefore, might be more responsive to that campaign. For this reason, Facebook deprecated geographic networks in October 2009. School and company networks are still available, but their security is better, since to join one of these networks, a user has to provide a valid e-mail address from that institution (e.g., a university e-mail address).

2.2 The MySpace Social Network

MySpace was the first social network to gain significant popularity among Internet users. The basic idea of this network is to provide each user with a web page, which the user can then personalize with information about herself and her interests. Even though MySpace has also the concept of "friendship," like Facebook, MySpace pages are public by default. Therefore, it is easier for a malicious user to obtain sensitive information about a user on MySpace than on Facebook. Users might be profiled by gender, age, or nationality, and an aimed spam campaign could target a specific group of users to enhance its effectiveness.

MySpace used to be the largest social network on the Internet. Although it is steadily losing users, who are mainly moving to Facebook [2], it remains the third most visited site of its kind on the Internet.

2.3 The Twitter Social Network

Twitter is a much simpler social network than Facebook and MySpace. It is designed as a microblogging platform, where users send short text messages (i.e., *tweets*) that appear on their friends' pages. Unlike Facebook and MySpace, no personal information is shown on Twitter pages by default. Users are identified only by a username and, optionally, by a real name. To profile a user, it is possible to analyze the tweets she sends, and the feeds to which she is subscribed. However, this is significantly more difficult than on the other social networks.

A Twitter user can start "following" another user. As a consequence, she receives the user's tweets on her own page. The user who is "followed" can, if she wants, follow the other one back. Tweets can be grouped by *hashtags*, which are popular words, beginning with a "#" character. This allows users to efficiently search who is posting topics of interest at a certain time. When a user likes someone's tweet, he can decide to *retweet* it. As a result, that message is shown to all her followers. By default, profiles on Twitter are public, but a user can decide to protect her profile. By doing that, anyone wanting to follow the user needs her permission. According to the same statistics, Twitter is the social network that has the fastest growing rate on the Internet. During the last year, it reported a 660% increase in visits [2].

2.4 Related Work

The success of social networks has attracted the attention of security researchers. Since social networks are strongly based on the notion of a network of trust, the exploitation of this trust might lead to significant consequences. In 2008, a Sophos experiment showed that 41% of the Facebook users who were contacted acknowledged a friend request from a random person [8]. Bilge et al. [10] show that after an attacker has entered the network of trust of a victim, the victim will likely click on any link contained in the messages posted, irrespective of whether she knows the attacker in real life or not. Another interesting finding was reported by Jagatic et al. [13]. The authors found that phishing attempts are more likely to succeed if the attacker uses stolen information from victims' friends in social networks to craft their phishing emails. There are also botnets that target social networks, such as koobface [9].

Brown et al. [12] showed how it would be possible for spammers to craft targeted spam by leveraging the information available in online social networks. As for Twitter, Krishnamurthy et al. studied the network, providing some characterization of Twitter users [14]. Yardi et al. [18] ran an experiment on Twitter spam. They created a popular hashtag on Twitter, and observed that spammers started using it in their messages. They also discuss some features that might allow one to distinguish a spammer from legitimate users, such as node degree and frequency of messages. Another work that studied social network spam using honey-profiles was conducted by Webb et al. in 2008 [17]. For this experiment, 51 profiles were created on MySpace, which was the largest social network at the time. The study showed a significant spam activity. The honey-profiles were contacted by 1,570 spam bots over a five-month period.

Compared to their work, our study is substantially larger in size and covers three major social networks, and the honeypot population we used is representative of the average population of these networks, both from an age and nationality point of view. Moreover, we leverage our observation to develop a system able to detect spammers on social networks.

This system has detected thousands of spam accounts on Twitter, which have been subsequently deleted.

3. DATA COLLECTION

The first goal of our paper was to understand the extent to which spam is a problem on social networks, as well as the characterization of spam activity. To this end, we created 900 profiles on Facebook, MySpace, and Twitter, 300 on each platform. The purpose of these accounts was to log the traffic (e.g., friend requests, messages, invitations) they receive from other users of the network. Due to the similarity of these profiles to honeypots [4], we call these accounts *honey-profiles*.

3.1 Honey-Profiles

Our goal was to create a number of honey-profiles that reflect a representative selection of the population of the social networks we analyzed. To this end, we first crawled each social network to collect common profile data.

On Facebook, we joined 16 geographic networks, using a small number of manually-created accounts. This was possible because, at the time, geographic networks were still available. Since we wanted to create profiles reflecting a diverse population, we joined networks on all continents (except Antarctica and Australia): the Los Angeles and New York networks for North America, the London, France, Italy, Germany, Russia, and Spain ones for Europe, the China, Japan, India, and Saudi Arabia ones for Asia, the Algeria and Nigeria ones for Africa, and the Brazil and Argentina networks for South America. For each network, we crawled 2,000 accounts at random, logging names, ages, and gender (which is the basic information required to create a profile on Facebook). Afterwards, we randomly mixed this data (names, surnames, and ages) and created the honey-profiles. Gender was determined by the first name. Each profile was assigned to a network. Accounts created using data from a certain network were assigned either to this network or to a network where the main language spoken was the same (e.g., profiles created from accounts in the France network were used in networks associated with francophone countries). This was a manual process. For larger networks (e.g., New York, Germany, Italy) up to three accounts were created, while only one account was set up for smaller ones. In total, we created 300 accounts on the Facebook platform.

On MySpace, we crawled 4,000 accounts in total. This was easier than on Facebook because, as mentioned in Section 2.2, most profile pages are public. Similar to Facebook, our aim was to generate "average" profiles based on the user population of the social network. After data collection, we looked for common names and ages from profiles with different languages, and created profiles in most nations of the world. We created 300 accounts on MySpace for our experiment.

While on Facebook and MySpace, birth date and gender are needed for registration, on Twitter, the only information required for signing up is a full name and a profile name. Therefore, we did not find it necessary to crawl the social network for "average" profile information, and we simply used first names and surnames from the other social networks. For each account, the profile name has been chosen as a concatenation of the first and last name, plus a random number to avoid conflicts with already existing accounts. Similarly to the other networks, we created 300 profiles.

We did not create more than 300 profiles on each network because registration is a semi-automated process. More precisely, even though we could automatically fill the forms required for registration, we still needed a human to solve the CAPTCHAs involved in the process.

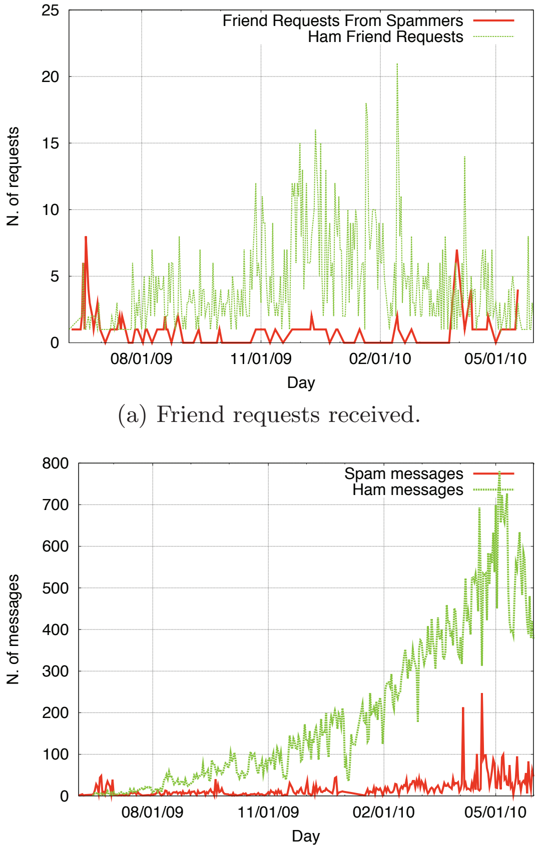
3.2 Collection of Data

After having created our honey-profiles, we ran scripts that periodically connected to those accounts and checked for activity. We decided that our accounts should act in a passive way. Therefore, we did not send any friend requests, but accepted all those that were received.

In a social network, the first action a malicious user would likely execute to get in touch with his victims is to send them a friend request. This might be done to attract the user to the spammer's profile to view the spam messages (on MySpace) or to invite her to accept the friendship and start seeing the spammer's messages in her own feed (on Facebook and Twitter).

After having acknowledged a request (i.e., accepted the friendship on Facebook and MySpace or started following the user on Twitter), we logged all the information needed to detect malicious activity. More precisely, we logged every email notification received from the social networks, as well as all the requests and messages seen on the honey-profiles. On some networks, such as Facebook, the notifications and messages might be of different types (e.g., application and group invitations, video posts, status messages, private messages), while on other platforms, they are more uniform (e.g., on Twitter, they are always short text messages). We logged all types of requests on Facebook, as well as wall posts, status updates, and private messages. On MySpace, we recorded mood updates, wall posts, and messages. On Twitter, we logged tweets and direct messages.

Our scripts ran continuously for 12 months for Facebook (from June 6, 2009 to June 6, 2010), and for 11 months for MySpace and Twitter (from June 24, 2009 to June 6, 2010), periodically visiting each account. The visits had



(a) Friend requests received.

(b) Messages received.

Figure 1: Activity observed on Facebook

to be performed slowly (approximately one account visited every 2 minutes) to avoid being detected as a bot by the social networking site and, therefore, having the accounts deleted.

4. ANALYSIS OF COLLECTED DATA

As mentioned previously, the first action that a spammer would likely execute is to send friend requests to her victims. Only a fraction of the contacted users will acknowledge a request, since they do not know the real-life person associated with the account used by the bot[1]. On Twitter, the concept of friendship is slightly different, but the modus operandi of the spammers is the same: they start following victims, hoping that they will follow them back, starting to receive the spam content. From the perspective of our analysis, friendships and mutual follow relationships are equivalent. When a user accepts one of the friend requests, she lets the spammer enter her network of trust. In practice, this action has a major consequence: The victim starts to see messages received from the spammer in her own news/message feed. This kind of spamming is very effective, because the spammer has only to write a single message (e.g., a status update on Facebook), and the message appears in the feeds of all the victims. Depending on the social network, the nature of these messages can change: they are *status updates* on Facebook, *status* or *mood updates* on MySpace, and *tweets* on Twitter.

Network	Overall	Spammers
Facebook	3,831	173
MySpace	22	8
Twitter	397	361

Table 1: Friend requests received on the various social networks.

Network	Overall	Spammers
Facebook	72,431	3,882
MySpace	25	0
Twitter	13,113	11,338

Table 2: Messages received on the various social networks.

During our study, we received a total of 4,250 friend requests. As can be seen in Table 1, the amount of requests varies from network to network. This might be caused by the different characteristics of the various social networks. As one would expect, we observed the largest amount of requests on Facebook, since it has the largest user base. Surprisingly, however, the majority of these requests proved not to come from spam bots, but from real users, looking for popularity or for real persons with the same name as one of our honey-profiles. Another surprising finding is that, on MySpace, we received a very low number of friend requests. It is not clear what is the reason of the disparity between this social network and Facebook, since MySpace also provides a mechanism to easily post messages on users' pages. Daily statistics for friend requests received on Facebook and Twitter are shown in Figures 1(a) and 2(a).

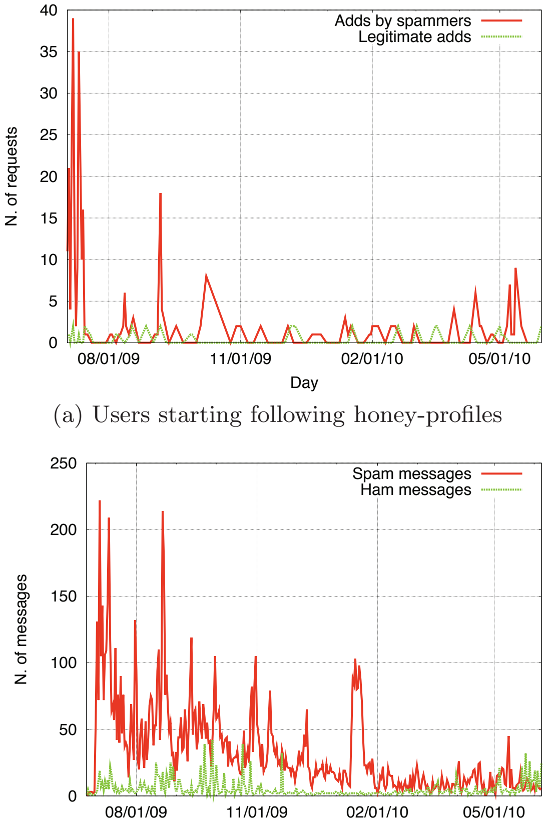
Information about the logged messages is shown in Table 2. Overall, we observed 85,569 messages. Again, there is a big disparity between the three social networks. On Twitter, interestingly, we recorded the largest amount of spam messages. Given the smaller size of the network's user base, this is surprising. Daily statistics for messages received on Facebook are shown in Figure 1(b), while those for Twitter are reported in Figure 2(b). We do not show a graph for MySpace because the number of messages we received was very low.

On Facebook, we also observed a fair amount of invitations to applications, groups, and events, as well as posting of photos and videos in our honey-profiles' feeds. However, since none of them were spam, we ignored them for the subsequent analysis.

4.1 Identification of Spam Accounts

Tables 1 and 2 show the breakdown of requests that were received by our honey-profiles. We can see that the honey-profiles did not only receive friend requests and messages from spammers, but also a surprising amount from legitimate accounts. Even if friend requests are unsolicited, they are not always the result of spammers who reach out. In particular, many social network users aim to increase their popularity by adding as friends people they do not know. On

[1] We assume that most spam accounts are managed in an automated fashion. Therefore, from this point on, we will use the terms spam profile and *bots* interchangeably.



(a) Users starting following honey-profiles

(b) Messages received

Figure 2: Activity observed on Twitter.

Facebook, since all our honey-profiles were members of a geographic network (as long as these were available), it is also possible that people looking for local "friends" would have contacted some of our accounts. In particular, we observed that this occurs with more frequency on smaller networks (in particular, some Middle Eastern and African ones). Moreover, since we picked random combinations of first and last names, it happened that some of our honey-profiles had the same name as a real person, and, as a consequence, the account was contacted by real friends of this person. Since not all friend requests and messages are malicious, we had to distinguish between spammers and benign users.

To discriminate between real users and spam bots, we started to manually check all the profiles that contacted us. During this process, we noticed that spam bots share some common traits, and formalized them in features that we then used for automated spam detection. We will describe these features in detail in Section 5.

We found that, of the original 3,831 accounts that contacted us on Facebook, 173 were spammers. Moreover, on Facebook, during the last months of logging, the ratio of spam messages compared to legitimate ones dramatically dropped. The reason is that when a legitimate user adds our honey-profile to her friend list, this honey-profile starts appearing on her friends' pages as a friend suggestion. This leads to a number of additional friend requests (and messages) from real users. On MySpace, we detected 8 spammers. On Twitter, we detected 361 spammers out of 397 contacts.

4.2 Spam Bot Analysis

The spam bots that we identified showed different levels of activity and different strategies to deliver spam. Based on their spam strategy, we distinguish four categories of bots:

1. **Displayer**: Bots that do not post spam messages, but only display some spam content on their own profile pages. In order to view spam content, a victim has to manually visit the profile page of the bot. This kind of bots is likely to be the least effective in terms of people reached. All the detected MySpace bots belonged to this category, as well as two Facebook bots.

2. **Bragger**: Bots that post messages to their own feed. These messages vary according to the networks: on Facebook, these messages are usually status updates, while on Twitter these are the tweets. The result of this action is that the spam message is distributed and shown on all the victims' feeds. However, the spam is not shown on the victim's profile when the page is visited by someone else (i.e., a victim's friends). Therefore, the spam campaign reaches only victims who are directly connected with the spam bot. 163 bots on Facebook belonged to this category, as well as 341 bots on Twitter.

3. **Poster**: Bots that send a direct message to each victim. This can be achieved in different ways, depending on the social network. On Facebook, for example, the message might be a post on a victim's wall. The spam is shown on the victims feed, but, unlike the case of a "bragger", can be viewed also by victim's friends visiting her profile page. This is the most effective way of spamming, because it reaches a greater number of users compared to the previous two. Eight bots from this category have been detected, all of them on the Facebook network. Koobface-related messages also belong to this category (see [9]).

4. **Whisperer**: Bots that send private messages to their victims. As for "poster" bots, these messages have to be addressed to a specific user. The difference, however, is that this time the victim is the only one seeing the spam message. This type of bots is fairly common on Twitter, where spam bots send direct messages to their victim. We observed 20 bots of this kind on this network, but none on Facebook and MySpace.

We then examined the activity of spam bots on different networks. On Facebook, we observed an average of 11 spam messages per day, while, on Twitter, the average number of messages observed was 34. On MySpace, we did not observe any direct spam message. The reason is that all the spam bots on MySpace are "displayers." The difference between Twitter and Facebook activity is caused by the apparently different responses of the two social networks to spam. More precisely, we observed that Facebook seems to be much more aggressive in fighting spam. This is demonstrated by the fact that, on Facebook, the average lifetime of a spam account was four days, while on Twitter, it was 31 days. On

MySpace, no spam accounts have been deleted during our observation.

As shown in Figures 1(a) and 2(a), many spam requests arrived during the first days of our experiment, especially on Facebook. All the early-days spammers have been quickly deleted from Facebook (the one with the longest life lasted one month), while most of the Twitter ones were deleted only after we flagged them to their spam team.

It is also interesting to look at the time of the day when messages and friend requests are sent. The reason is that bots might get activated periodically or at specific times to send their messages. Benign activity, on the other hand, follows the natural diurnal pattern. During our observation, we noticed that some bots showed a higher activity around midnight (GMT -7), while in the same period of time, the ham messages registered a low.

Another way to study the effectiveness of spam activity is to look at how many users acknowledged friend requests on the different networks. On Facebook, the average number of confirmed friends of spam bots is 21, on MySpace it is 31, while on Twitter, it is 350. We assume that the difference in number of people reached is probably due to the different lifetime of the bots in the different networks. The low activity of the bots on MySpace might be the cause of both the low numbers of bots detected on that network and their longer lifetime.

We identified two kinds of bot behavior: stealthy and greedy bots. Greedy ones include a spam content in every message they send. They are easier to detect, and might lead users to flag bots as spammers or to revoke their friendship status. Stealthy bots, on the other hand, send messages that look legitimate, and only once in a while inject a malicious message. Since they look like legitimate profiles, they might convince more people to accept and maintain friendships.

Of the 534 spam bots detected, 416 were greedy and 98 were stealthy (note that ten spam profiles were "displayers," and 20 were "whisperers." These bots, therefore, did not use updates or tweets to spam).

Another interesting observation is that spam bots are usually less active than legitimate users. This probably happens because sending out too many messages would make detection by the social network too easy. For this reason, most spam profiles we observed, both on Facebook and Twitter, sent less than 20 messages during their life span.

While observing Facebook spammers, we also noticed that many of them did not seem to pick victims randomly, but, instead, they seemed to follow certain criteria. In particular, most of their victims happened to be male. This was particularly true for campaigns advertising adult websites. Since Facebook does not provide an easy way to search for people based on gender, the only way spammers can identify their victims is by looking for male first names. This intuition led us to another observation. The list of victims targeted by these bots usually shows an anomalous repetition of people with the same first name (e.g., tens of profiles with only four different given names). This might happen because spam bots are given lists of first names to target. In addition, Facebook people search does not make a difference between first and last name while searching. For this reason, these gender-aware bots sometimes targeted female users who happened to have a male name as last name (e.g., Wayne).

Mobile Interface.

Most social networking sites have introduced techniques to prevent automatic account generation and message sending. On Facebook, for example, a user is required to solve a CAPTCHA [5] every time she tries to send a friend request. A CAPTCHA has to be solved also every time an account is created. Moreover, the site uses a very complicated JavaScript environment that makes it difficult for bots to interact with the pages. On the other hand, the complexity of these sites made them not very attractive to mobile Internet users, who use less powerful devices and slower connections.

To attract more users and to make their platform more accessible from any kind of device, major social networks launched mobile versions of their sites. These versions offer the main functionality of the complete social networking sites, but in a simpler fashion. To improve usability, no JavaScript is present on these pages, and no CAPTCHAs are required to send friend requests. This has made social networks more accessible from everywhere. However, the mobile environment provides spammers with an easy way to interact with these sites and carry out their tasks. This is confirmed by our analysis: 80% of bots we detected on Facebook used the mobile site to send their spam messages. However, to create an account, it is still necessary to go through the non-mobile version of the site. For Twitter spam, there is no need for the bots to use the mobile site, since an API to interact with the network is provided, and, in any case, there is no need to solve CAPTCHAs other than the one needed to create a profile.

5. SPAM PROFILE DETECTION

Based on our understanding of spam activity in social networks, the next goal was to leverage these insights to develop techniques to detect spammers in the wild. We decided to focus on detecting "bragger" and "poster" spammers, since they do not require real profiles for detection, but are just detectable by looking at their feeds. We used machine learning techniques to classify spammers and legitimate users. To detect whether a given profile belongs to a spammer or not, we developed six features, which are:

FF ratio (R): The first feature compares the number of friend requests that a user sent to the number of friends she has. Since a bot is not a real person, and, therefore, nobody knows him/her in real life, only a fraction of the profiles contacted would acknowledge a friend request. Thus, one would expect a distinct difference between the number of friend requests sent and the number of those that are acknowledged. More precisely, we expect the ratio of friend requests to actual friends to be large for spammers and low for regular users. Unfortunately, the number of friend requests sent is not public on Facebook and on MySpace. On Twitter, on the other hand, the number of users a profile started to follow is public. Therefore, we can compute the ratio $R = \textit{following} / \textit{followers}$ (where following, in the Twitter jargon, is the number of friend requests sent, and followers is the number of users who accepted the request).

URL ratio (U): The second feature to detect a bot is the presence of URLs in the logged messages. To attract users to spam web pages, bots are likely to send URLs in their messages. Therefore, we introduce the ratio U as:

$$U = \textit{messages\_containing\_urls} \ / \ \textit{total\_messages}.$$

Since, in the case of Facebook, most messages with URLs (link and video share, group invitations) contain a URL to other Facebook pages, we only count URLs pointing to a third party site when computing this feature.

Message Similarity (S): The third feature consists in leveraging the similarity among the messages sent by a user. Most bots we observed sent very similar messages, considering both message size and content, as well as the advertised sites. Of course, on Twitter, where the maximum size of the messages is 140 characters, message similarity is less significant than on Facebook and MySpace, where we logged messages up to 1,100 characters. We introduced the similarity parameter S, which is defined as follows:

$$S = \frac{\sum_{p \in P} c(p)}{l_a l_p},$$

where P is the set of possible message-to-message combinations among any two messages logged for a certain account, p is a single pair, $c(p)$ is a function calculating the number of words two messages share, l_a is the average length of messages posted by that user, and l_p is the number of message combinations. The idea behind this formula is that a profile sending similar messages will have a low value of S.

Friend Choice (F): The fourth feature attempts to detect whether a profile likely used a list of names to pick its friends or not. We call this feature F, and we define it as:

$$F = \frac{T_n}{D_n},$$

where T_n is the total number of names among the profiles' friend, and D_n is the number of distinct first names. Our observation showed that legitimate profiles have values of this feature that are close to 1, while spammers might reach values of 2 or more.

Messages Sent (M): We use the number of messages sent by a profile as a feature. This is based on the observation that profiles that send out hundreds of messages are less likely to be spammers, given that, in our initial analysis, most spam bots sent less that 20 messages.

Friend Number (FN): Finally we look at the number of friends a profile has. The idea is that profiles with thousands of friends are less likely to be spammers that the ones with a few.

Given our general set of features, we built two systems to detect spam bots on Facebook and Twitter. Since there are differences between these two social networks, some features had to be slightly modified to fit the characteristics of the particular social network. However, the general approach remains the same. We used the Weka framework [7] with a Random Forest algorithm [11] for our classifier. We chose this algorithm because it was the one that gave the best accuracy and lowest false positive ratio when we performed the cross-validation of the training set.

5.1 Spam Detection on Facebook

The main issue when analyzing Facebook is to obtain a suitable amount of data to analyze. Most profiles are private, and only their friends can see their walls. At the beginning of this study, geographic networks were still available, but they were discontinued in October 2009. Therefore, we used data from various geographic networks, crawled between April 28 and July 8 2009, to test our approach.

Since on Facebook the number of friend requests sent out is not public, we could not apply the R feature.

We trained our classifier using 1,000 profiles. We used the 173 spam bots that contacted our honey-profiles as samples for spammers, and 827 manually checked profiles from the Los Angeles network as samples for legitimate users. A 10-fold cross validation on this training data set yielded an estimated false positive ratio of 2% and a false negative ratio of 1%. We then applied our classifier to 790,951 profiles, belonging to the Los Angeles and New York networks. We detected 130 spammers in this dataset. Among these, 7 were false positives. The reason for this low number of detected spammers might be that spam bots typically do not join geographic networks. This hypothesis is corroborated by the fact that among the spam profiles that contacted out honey profiles, none was a member of a geographic network. We then randomly picked 100 profiles, classified as legitimate. We manually looked at them to search for false negatives. None of them turned out to be a spammer.

5.2 Spam Detection on Twitter

On Twitter, is much easier to obtain data than on Facebook, since most profiles are public. This gave us the possibility to develop a system that is able to detect spammers in the wild. The results of our analysis were then sent to Twitter, who verified that the accounts were indeed sending spam and removed them.

To train our classifier, we picked 500 spam profiles, coming either from the ones that contacted our honey profiles, or manually selected from the public timeline. We included profiles from the public timeline to increase diversity among spam profiles in our training dataset. Among the profiles from the public timeline, we chose the ones that stood out from the average for at least one of the R, U, and S features. We also picked 500 legitimate profiles from the public timeline. This was a manual process, to make sure that no spammers were miscategorized in the training set. The R feature was modified to reflect the number of followers a profile has. This was done because legitimate profiles with a fairly high number of followers (e.g., 300), but following thousands of other profiles, have a high value of R. This is a typical situation for legitimate accounts following news profiles, and would have led to false positives in our system. Therefore, we defined a new feature R', which is the R value divided by the number of followers a profile has. We used it instead of R for our classification.

After having trained the classifier, it was clear that the F feature is not useful to detect spammers on Twitter, since both spammers and legitimate profiles in the training set had very similar values for this parameter. This suggests that Twitter spam bots do not pick their victims based on their name. Therefore, we removed the F feature from the Twitter spam classifier. A 10-fold cross validation for the classifier with the updated feature set yielded an estimated false positive ratio of 2.5% and a false negative ratio of 3% on the training set.

Given the promising results on the training set and the possibility to access most profiles, we decided to use our classifier to detect spammers in real time on Twitter. The main problem we faced while building our system was the crawling speed. Twitter limited our machine to execute only 20,000 API calls per hour. Thus, to avoid wasting our limited API calls, we executed Google searches for the most common words in tweets sent by the already detected spammers, and we crawled those profiles that were returned as results. This approach has the problem that we can only detect profiles that send tweets similar to those of previously observed bots. To address this limitation, we created a public service where Twitter users can flag profiles as spammers. After a user has flagged someone as a spammer, we run our classifier on this profile data. If the profile is detected as a spammer, we add this profile to our detected spam set, enabling our system to find other profiles that sent out similar tweets.

Every time we detected a spam profile, we submitted it to Twitter. During a period of three months, from March 06, 2010 to June 06, 2010, we crawled 135,834 profiles, detecting 15,932 of those as spammers. We sent this list of profiles to Twitter, and only 75 were reported by them to be false positives. All the other submitted profiles were deleted. In order to evaluate the false negative ratio, we randomly picked 100 profiles, classified as legitimate by our system. We then manually checked at them, finding out that 6 were false negatives.

To show that our targeted crawling does not affect our accuracy or false positive ratio, but just narrowed down the set of profiles to crawl, we picked 40,000 profiles at random from the public timeline and crawled them. Among these, we detected 102 spammers, with a single false positive. We can see that our crawling is effective, since the percentage of spammers in our targeted (crawled) dataset is 11%, whereas in the random set, it is 0.25%. On the other hand, the false positive ratio on in both datasets is similarly low.

5.3 Identification of Spam Campaigns

After having identified single spammers, we analyzed the data to identify larger-scale spam campaigns. With "spam campaign," we refer to multiple spam profiles that act under the coordination of a single spammer. We consider two bots posting messages with URLs pointing to the same site as being part of the same campaign. Most bots hide the real URL that their links are pointing to by using URL shortening services (for example, *tinyurl* [6]). This is typically done to avoid easy detection by social networks administrators and by the users, as well as to meet the message length requirements of some platforms (in particular, Twitter). To determine the actual site that a shortened URL points to, we visited all the URLs that we observed. Then, we clustered all the profiles that advertised the same page. We list the top eight campaigns, based on the number of observed messages, in Table 3. Since we had most detections on Twitter, these campaigns targeted that network. It is interesting to notice, however, that bots belonging to three of them were observed on Facebook as well.

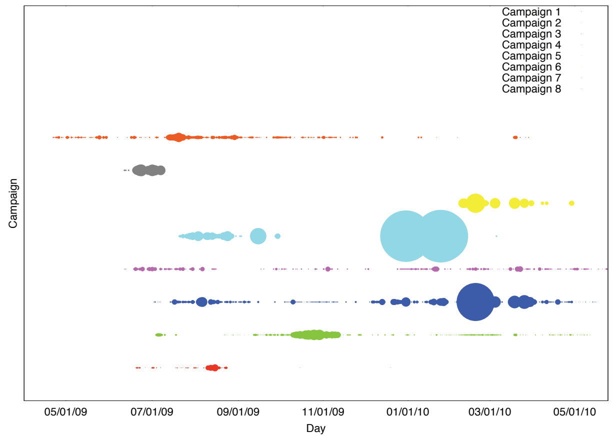


Figure 3: Activity of campaigns over time.

Some campaigns showed a large number of bots, each sending a few messages per day, while others send many messages using few bots. In addition, the fact that bots of a campaign can act in a stealthy or greedy way (see Section 4.2) leads to significantly different outcomes. Greedy bots that send spam with each message are easier to detect by the social network administrators. On the other hand, a low-traffic spam campaign is not easy to detect. For example, the bots from Campaign 1 sent 0.79 messages per day, while the bots from the second campaign sent 0.08 messages per day on average. The result was that the bots from Campaign 1 have an average lifetime of 25 days, while the bots of Campaign 2 lasted 135 days on average. In addition, Campaign 2 reached more victims, as shown by an average of 94 friends (victims) per bot, while Campaign 1 only reached 52. This suggests that a relationship exists between the lifetime of bots and the number of victims targeted. Clearly, an effective campaign should be able to reach many users, and having bots that live longer might be a good way to achieve this objective.

From the point of view of victims reached, stealthy campaigns are more effective. Campaigns 4 and 7 both used a stealthy approach. Of the messages sent, only 20-40% contained spam content. As a result, bots from Campaign 4 had an average lifetime of 120 days, and started following 460 profiles each. Among these, 87 users on average followed the bots back. Campaign 7 was the most effective among Twitter campaigns, both considering the number of victims and the average bot lifetime. To achieve this, this campaign combined a low rate of messages per day with a stealth way of operating. The bots in this campaign have an average lifetime of 198 days and 1,787 victims, of which, on average, 112 acknowledged the friend request.

From the observations of the various campaigns, we developed a metric that allows us to predict the success of a campaign. We consider a campaign successful if the bots belonging to it have a long lifetime. For this metric, we introduce the parameter G_c, defined as follows:

$$G_c = \frac{M_d^{-1} \cdot S_d}{((\sqrt{M_d^{-1}} \cdot S_d)+1)^2}, \; 0 \leq G_c \leq 1.$$

In the above formula, M_d is the average number of messages per day sent and S_d is the ratio of actual spam messages ($0 \leq S_d \leq 1$). Empirically, we see that campaigns with a value of G_c close to 1 have a long lifetime (for example, Campaign 7 has $G_c = 0.88$, while Campaign 2 has $G_c = 0.60$), while for campaigns with a lower value of this parameter, the average lifetime decreases significantly (Campaign 1 has $G_c = 0.28$ and Campaign 5 has $G_c = 0.16$). Thus, we can infer that a value of 0.5 or higher for G_c indicates that a campaign has a good chance to be successful. Of course, if a campaign is active for some time, a social network might develop other means to detect spam bots belonging to it (e.g., a blacklist of the URLs included in the messages).

Activity of bots from different campaigns is shown in Figure 3. Each row represents a campaign. For each day in

#	SN	Bots	# Mes.	Mes./day	Avg. vic.	Avg. lif.	G_c	Site adv.
1	T	485	1,020	0.79	52	25	0.28	Adult Dating
2	T	282	9,343	0.08	94	135	0.60	Ad Network
3	T,F	2,430	28,607	0.32	36	52	0.42	Adult Dating
4	T	137	3,213	0.15	87	120	0.56	Making Money
5	T,F	5,530	83,550	1.88	18	8	0.16	Adult Site
6	T,F	687	7,298	1.67	23	10	0.18	Adult Dating
7	T	860	4,929	0.05	112	198	0.88	Making Money
8	T	103	5,448	0.4	43	33	0.37	Ad Network

Table 3: Spam campaigns observed.

which we observed some activity from that campaign, a circle is drawn. The size of circles varies according to the number of messages observed that day. As can be seen, some campaigns have been active over the entire period of the study, while some have not been so successful.

We then tried to understand how bots choose their victims. The behavior seems not to be uniform for the various campaigns. For example, we noticed that many victims of Campaign 2 shared the same hashtag (e.g., "#iloveitwhen") in their tweets. Bots might have been crawling for people sending messages with such tag, and started following them. On the other hand, we noticed that Campaigns 4 and 5 targeted an anomalous number of private profiles. Looking at their victims, 12% of them had a private profile, while for a random picked set of 1,000 users from the public timeline, this ratio was 4%. This suggests that bots from these campaigns did not crawl any timeline, since tweets from users with a private profile do not appear on them.

6. CONCLUSIONS

Social networking sites have millions of users from all over the world. The ease of reaching these users, as well as the possibility to take advantage of the information stored in their profiles, attracts spammers and other malicious users.

In this paper, we showed that spam on social networks is a problem. For our study, we created a population of 900 honey-profiles on three major social networks and observed the traffic they received. We then developed techniques to identify single spam bots, as well as large-scale campaigns. We also showed how our techniques help to detect spam profiles even when they do not contact a honey-profile. We believe that these techniques can help social networks to improve their security and detect malicious users. In fact, we develop a tool to detect spammers on Twitter. Providing Twitter the results of our analysis thousands of spamming accounts were shut down.

7. ACKNOWLEDGMENTS

This work was supported by the ONR under grant N000140911042 and by the National Science Foundation (NSF) under grants CNS-0845559 and CNS-0905537.

8. REFERENCES

[1] Alexa top 500 global sites. `http://www.alexa.com/topsites`.

[2] Compete site comparison. `http://siteanalytics.compete.com/facebook.com+myspace.com+twitter.com/`.

[3] Facebook statistics. `http://www.facebook.com/press/info.php?statistics`.

[4] Honeypots. `http://en.wikipedia.org/wiki/Honeypot\_computing`.

[5] The recaptcha project. `http://recaptcha.net/`.

[6] Tinyurl. `http://tinyurl.com/`.

[7] Weka - data mining open source program. `http://www.cs.waikato.ac.nz/ml/weka/`.

[8] Sophos facebook id probe. `http://www.sophos.com/pressoffice/news/articles/2007/08/facebook.html`, 2008.

[9] J. Baltazar, J. Costoya, and R. Flores. Koobface: The largest web 2.0 botnet explained. 2009.

[10] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your contacts are belong to us: Automated identity theft attacks on social networks. In *World Wide Web Conference*, 2009.

[11] L. Breiman. Random forests. In *Machine Learning*, 2001.

[12] G. Brown, T. Howe, M. Ihbe, A. Prakash, and K. Borders. Social networks and context-aware spam. In *ACM Conference on Supportive Cooperative Work*, 2008.

[13] T.N. Jagatic, N.A. Johnson, M. Jakobsson, and T.N. Jagatif. Social phishing. *Comm. ACM*, 50(10):94–100, 2007.

[14] B. Krishnamurthy, P. Gill, , and M. Aritt. A few chirps about twitter. In *USENIX Workshop on Online Social Networks*, 2008.

[15] S. Moyer and N. Hamiel. Satan is on my friends list: Attacking social networks. `http://www.blackhat.com/html/bh-usa-08/bh-usa-08-archive.html`, 2008.

[16] Harris Interactive Public Relations Research. A study of social networks scams. 2008.

[17] S. Webb, J. Caverlee, , and C.Pu. Social honeypots: Making friends with a spammer near you. In *Conference on Email and Anti-Spam (CEAS 2008)*, 2008.

[18] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd. Detecting spam in a twitter network. *First Monday*, 15(1), 2010.