Detecting Spammers on Social Networks

Gianluca Stringhini, Christopher Kruegel, Giovanni Vigna

University of California, Santa Barbara
26th ACSAC(December, 2010)

Presented by
Alankrit Chona 2009CS10176
Nikita Gupta 2009CS50248
Outline

- Motivation
- The Popular Social Networks
- Data Collection
- Analysis of Collected Data
- Spam Profile Detection
- Our Proposals
Idea

- To conduct a comprehensive survey of existing and potential attack behaviors in social network sites
- Identify patterns in such attack behaviors
- Detect Spammers in a real-world social network
- Review existing solutions, measurement as well as defense mechanisms
What is SPAM?

- “Unsolicited usually commercial e-mail sent to a large number of addresses” – Merriam Webster Online

- As the Internet has supported new applications, many other forms are common, requiring a much broader definition

Capturing user attention unjustifiably in Internet enabled applications (e-mail, Web, Social Media etc..)
SPAM TAXONOMY

INTERNET SPAM

DIRECT
- E-Mail Spam
- IM Spam (SPIM)
- Social Network Spam

INDIRECT
- Spam Blogs (Splogs)
- General Web Spam

[Forms]

[Mechanisms]

Social Media Spam

Spamdexing
Auto-generated and/or Plagiarized Content

Advertisements in Profitable Contexts

Link Farms to promote other spam pages
Sites such as Facebook, MySpace, and Twitter are consistently among the top 20 most-viewed web sites of the Internet.

In 2008, 83% of the users of social networks have received at least one unwanted friend request or message.

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.
The Popular Social Networks
Facebook administrators claim to have more than 400 million active users all over the world.

Most Facebook users were grouped in networks, where people coming from a certain country, town, or school could find their neighbours or peers.

Default privacy setting was to allow all people in the same network to view each other’s profiles.
- Facebook deprecated geographic networks in October 2009.
You CLICK!
Is it same as real life?

Facebook

Your wall

Hey afsana check this embarrassing video Blame.

[VIDEO] Yeah! It happens on Live Television!
play-all-now.blogspot.com
Lo! Checkout this video its a very embarrassing moment for her. This shocking...

That last afsana check the sad post I dare you can watch this.

[Video] Girl killed herself after her dad posted a secret of her on here fb wall!!!
smj3kdxh.blogspot.com
click here to see dad post and enact suicide letter, you will really be shocked... IP. This cool....

You click!

It spreads as a post to your added friends.
An extension to our web browser which accesses the websites we visit and posts such content on our behalf

You click!

It spreads as a post to your added friends
It is designed as a microblogging platform, where users send short text messages (i.e., *tweets*) that appear on their friends’ pages.

Users are identified only by a username and, optionally, by a real name.

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.
900 profiles on Facebook, MySpace, and Twitter, 300 created on each platform

A honeypot is a "trap" set to detect, deflect, or in some manner counteract attempts at unauthorized use of information systems

Due to the similarity of these profiles to honeypots, these accounts are called as honey-profiles
## Honey-Profiles

### Mail Center

### Friend Request Manager

<table>
<thead>
<tr>
<th>Date:</th>
<th>From:</th>
<th>Confirmation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 Dec 2007 11:46 AM</td>
<td><img src="image" alt="Profile Picture" /></td>
<td>Delia wants to be your friend!</td>
</tr>
<tr>
<td></td>
<td><img src="image" alt="Online Now!" /></td>
<td><img src="image" alt="Approve" /> <img src="image" alt="Deny" /> <img src="image" alt="Spam" /></td>
</tr>
</tbody>
</table>

Send Message

---

**Check/Uncheck All**

![Approve](image) ![Deny](image) ![Spam](image)
Honey–Profiles

Facebook

- Joined 16 geographic networks, using a small number of manually-created accounts
- Crawled 2,000 accounts at random, logging names, ages, and gender for each network
- Randomly mixed this data (names, surnames, and ages) and created the honey–profiles
Honey-Profiles

- MySpace
  - Crawled 4,000 accounts in total

- Twitter
  - Only information required for signing up is a full name and a profile name
Collection of Data

- No friend requests were sent, but all those that were received were accepted

- Logged every email notification received from the social networks, as well as all the requests and messages seen on the honey-profiles

- 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)
Collected Data

Looking for “local” friends

<table>
<thead>
<tr>
<th>Network</th>
<th>Overall</th>
<th>Spammers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>3,831</td>
<td>173</td>
</tr>
<tr>
<td>MySpace</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>Twitter</td>
<td>397</td>
<td>361</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Network</th>
<th>Overall</th>
<th>Spammers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>72,431</td>
<td>3,882</td>
</tr>
<tr>
<td>MySpace</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>Twitter</td>
<td>13,113</td>
<td>11,338</td>
</tr>
</tbody>
</table>

Table 2: Messages received on the various social networks.
Analysis of Collected Data

Facebook
Twitter
(a) Friend requests received.

(b) Messages received.
(a) Users starting following honey-profiles

(b) Messages received
Results

- Honey-profiles did not only receive friend requests and messages from spammers, but also a surprising amount from legitimate accounts.
  - In particular, many social network users aim to increase their popularity by adding as friends people they do not know.
- On Facebook, since all honey-profiles were members of a geographic network it may be that people looking for local “friends” would have contacted some of our accounts.
Spam Bot Analysis

Displayer
Bragger
Poster
Whisperer
Displayer

- Bots that only display some spam content on their own profile pages.

- All the detected MySpace bots belonged to this category.
Bots that post messages to their own feed

Messages vary according to the networks
- Facebook: status updates
- Twitter: tweets

Spam message is distributed and shown on all the victims’ feeds

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
Bragger Example
Bots that send a direct message to each victim

This can be achieved in different ways
- On Facebook, for example, the message might be a post on a victim’s wall

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.
Bots that send private messages to their victims

As for “poster” bots, these messages have to be addressed to a specific user

- Difference: Victim is the only one seeing the spam message.

Fairly common on Twitter, where spam bots send direct messages to their victim.
## Spam Bot Analysis

<table>
<thead>
<tr>
<th></th>
<th>Facebook</th>
<th>MySpace</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Displayer</td>
<td>2</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Bragger</td>
<td>163</td>
<td>0</td>
<td>341</td>
</tr>
<tr>
<td>Poster</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Whisperer</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>
Two kinds of bot behavior were identified
  ◦ Greedy: Include a spam content in every message they send (416)
  ◦ Stealthy: Send messages that look legitimate, and only once in a while inject a malicious message (98)

Most spam profiles under observation, both on Facebook and Twitter, sent less than 20 messages during their life span.
Used machine learning techniques to classify spammers and legitimate users.

Developed six features to detect whether a given profile belongs to a spammer or not.
Spam Profile Detection

- **FF ratio (R)**
  - The feature compares the number of friend requests that a user sent to the number of friends she has
    - Unfortunately, the number of friend requests sent is not public on Facebook and on MySpace
    - On Twitter, the number of users a profile started to follow is public.
  - \( R = \frac{\text{following}}{\text{followers}} \)
Spam Profile Detection

- URL ratio ($U$)
  - The feature to detect a bot is the presence of URLs in the logged messages
    - $U = \frac{\text{messages containing urls}}{\text{total messages}}$
  - Only count URLs pointing to a third party site when computing this feature
Spam Profile Detection

- Message Similarity ($S$)
  \[ S = \frac{\sum_{p \in P} c(p)}{l_a l_p} \]

- Friend Choice ($F$)
  \[ F = \frac{T_n}{D_n} \]
Spam Profile Detection

- Messages Sent (M)
- Friend Number (FN)
- The Weka framework with a Random Forest algorithm was used for the classifier.
Challenges
- Obtain a suitable amount of data to analyze
- Most profiles are private, and only their friends can see their walls
- Geographic networks discontinued in October 2009
- Used data from various geographic networks, crawled between April 28 and July 8 2009, to test our approach.

Modifications
- R feature not applicable
Trained classifier using 1,000 profiles
  ◦ 173 spam bots that contacted honey-profiles
  ◦ 827 manually checked profiles

From 790,951 profiles
  ◦ Detected: 130
  ◦ False positive: 7

From 100 profiles
  ◦ False negative: 0

Low number may be due to the fact that spam bots typically do not join geographic networks
Detection on Twitter

- Easier to obtain data than Facebook
  - Most profiles are public
- 500 spam profiles picked
- 500 legitimate profiles also picked

Modifications:
- Legitimate profiles with a fairly high number of followers (e.g., 300), but following thousands of other profiles, have a high value of R.
- Therefore, a new feature $R'$
  - \((R \text{ value})/(\text{the number of followers a profile has})\)
The main problem faced while building the system was the crawling speed

- Twitter limited the machine to execute only 20,000 API calls per hour

From March 06, 2010 to June 06, 2010, we crawled 135,834 profiles, detecting 15,932 of those as spammers.

- False positive: 75
Identification of Spam Campaigns

- Two bots posting messages with URLs pointing to the same site are considered part of the same campaign.
A relationship exists between the lifetime of bots and the number of victims targeted.
**Spam Campaigns—Results**

4,7 use a Stealthy Approach

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

A relationship exists between the lifetime of bots and the number of victims targeted
Success of a Campaign

- $G_c \geq 0.5$ : high success probability

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

Md: avg # of sent Mes/day

Sd: Ratio of actual spam msg
Conclusion

- Social networking sites attracts spammers
  - ease of reaching these users
  - possibility to take advantage of the information stored in their profiles
- Created a population of 900 honey-profiles on three major social networks and observed the traffic they received
- Developed techniques to identify single spam bots, as well as large-scale campaigns.
- These techniques can help social networks to improve their security and detect malicious users.
Our Proposals
Reverse Social Engineering Attacks

- Attacker does not initiate contact with victim
- Victim is tricked into contacting the attacker
- Techniques of abuse of friend-finding features
Reverse Social Engineering Attacks

- Once a reverse social engineering attack is successful, a wide range of attacks such as persuading victims to click on malicious links, blackmailing, identity theft, and phishing can be done!

- Can bypass current behavioral and filter-based detection techniques that aim to prevent wide-spread unsolicited contact.

- Third, if the victim contacts the attacker, less suspicion is raised, and there is a higher probability that a social engineering attack (e.g., phishing, a financial scam, information theft, etc.) will be successful.
References
