Detecting Facebook Synthetic Accounts
Gil Raytan and Ron Galay
Under the supervision and guidance of Amichai Shulman
Project Definition:

Creating heuristic algorithms for finding synthetic accounts on Facebook.

Why are we interested in synthetic accounts in the first place?
Why are we interested in synthetic accounts in the first place?

- Facebook can be used as a platform for spreading malware
- The vast majority of "foul play" is directed to advertisement and general spamming.
  - In order to use Facebook as a marketing tool, the marketer needs to gather a vast dataset of real Facebook accounts (meaning accounts that are used by real people), which he can address.

Synthetic accounts in control of an attacker create a "volume" of people, the sheer amount of them is sufficient to cause damage.
First we have explored the Facebook API

Acquiring and processing data:

- In the second section we explored different methodologies to gather information on users, and ultimately devised algorithms, that can conclude if a given account is synthetic.
- This part was divided to two sections as well. In the first we acquired a large base of suspected accounts, and in the second examined each account by measurement we decided on.

Analyzing of the results.
Relevant Information About Facebook

Facebook Users

Facebook Pages
We have used the following tools working on this project:

- **Using external application in order to gain an offline authentication key.**
- **We used the Python SDK to connect to the Facebook Graph API.**
- **We used FQL queries to retrieve efficiently information from the Facebook Graph API.**
- **The information we extracted was maintained in text files coded in UTF-8.**
- **We established a SQL -Server, and built the following tables: main_db, users_like_table, users_post_table. All the data was processed and inserted into these tables.**
Methodologies For Detecting Synthetic Accounts - Fan pages (1)

- **How is it related to our initial motivation?**
  - Fan pages can be hijacked for spamming:
    - posting comments on the wall or a given post
    - posting pictures that contain a spamming description
    - posting content which contains malicious software

- **We examined a fan page as a spammer would, and retrieved any allowed information.**
Methodologies For Detecting - Synthetic Accounts

Fan pages (2)

Algorithm description:
we have explored the 15 most popular fan pages on Facebook:

- Facebook (~ 35 millions fans)
- Coca Cola
- Justin Bieber
- Katy Perry
- Lady Gage
- Michael Jackson
- Megan Fox
- Rihanna
- South Park
- Texas HoldEm
- Xbox
- Youtube
- Adam Sandler
- FC Barcelona
- Barack Obama
Methodologies For Detecting - Synthetic Accounts

Fan pages (3)

Algorithm description:
for each fan page,
we have gathered all the relevant information:

- Wall posts
- Comments on wall posts
- Likes
- Comments and likes on photos posted on the page
- Comments and likes on videos posted on the page

Each block of data was saved with the name and the id of the user who posted the data.

The file containing the retrieved information was used as our repository of suspected users.
Methodologies For Detecting - Synthetic Accounts

Fan pages (4)

Algorithm description:

For each user kept in the suspected repository, we extracted any available relevant data.

- personal information (name, location, gender etc).
- personal news feed (including his statuses and friends posts on his wall)
- pages the user like
Methodologies For Detecting - Synthetic Accounts

Fan pages (5)

Classification Algorithm:

We have determined whether a suspected user is synthetic or not according to some parameters:

- comparing user posts with common spam words ("for free", "click here", urls etc).
- inspecting whether the same identical post appears in different locations (posts of user on fan page or on his wall).
- inspecting whether the user had a big mass of interaction with the fan pages.
- inspecting the pages the user like.
Methodologies For Detecting - Synthetic Accounts

Fan pages (6)

Results:

We have retrieved and processed 40,912 blocks of data from the fan pages during 2 full days of execution. The information we retrieved contains 11,142 unique Facebook users.

2,501 users were found synthetic according to the classification parameters we determined.
Examples -

users detected as synthetic (spam)

On facebook page
user_id 1175485735
"Earn real money for taking surveys and completing free offers.
Get started and receive $1.00 bonus now!
*********
http://apps.facebook.com/easycash-usd/?rfid=1175485735"

On ladygaga page
user_id 528069125
"HEY IF YOU WANT TO WIN A $20 TOASTER ADD THIS BAND THEY KICK ASS

On JustinBieber page
user_id 136013939794429
"Charlie Sheen Talking about Cocaine and Hookers http://www.youtube.com/watch?v=l9PHkyxrAzA"

On JustinBieber page
user_id 182458435123614
"Please visit my profile and click http://facebook.com/livingadvertisement I like it! no this post! click LIKE in my profile! This is my idea for a better life! Can you help me by clicking like it :-) Thank you!"
Example - detected users who like enormous number of pages

<table>
<thead>
<tr>
<th>USER ID</th>
<th>NUMBER OF LIKES</th>
</tr>
</thead>
<tbody>
<tr>
<td>100000933523635</td>
<td>146</td>
</tr>
<tr>
<td>100001250861472</td>
<td>247</td>
</tr>
<tr>
<td>100001369321729</td>
<td>3634</td>
</tr>
<tr>
<td>1056162042</td>
<td>231</td>
</tr>
<tr>
<td>569412096</td>
<td>106</td>
</tr>
<tr>
<td>864825073</td>
<td>368</td>
</tr>
<tr>
<td>1639000014</td>
<td>649</td>
</tr>
<tr>
<td>218471</td>
<td>187</td>
</tr>
<tr>
<td>2901279</td>
<td>309</td>
</tr>
<tr>
<td>514451010</td>
<td>725</td>
</tr>
<tr>
<td>585597475</td>
<td>263</td>
</tr>
</tbody>
</table>
Methodologies For Detecting
- Synthetic Accounts
Distribution of popular names in the USA (1)

- **The Facebook directory and it's relation to "black Facebook"**
- **USA as the center of internet technologies and trends**
- **How is it related to our initial motivation?**
  - **Those who really don't want to be found...**
Algorithm Description:

We used the U.S. Census Bureau statistics about names distribution among the USA to retrieve information about the users named with the most popular names.

For each user, we checked whether he is kept at the Facebook directory.
Methodologies For Detecting Synthetic Accounts - Distribution of popular names in the USA (3)

This methodology did not work exactly as we expected.

- There is a problem to retrieve information about "hidden" users by their names (possible by id)

- The popular names in the Facebook directory and in U.S. Census Bureau statistics were almost the same rates, and therefore we could not recognize "suspected names"

used for some minor FQL queries for fetching data on random users (random ids).
Further Development

Steps which can be done for expanding and improving the project:

- Inspecting users photos
- Code optimization
- Use FQL to retrieve random users not kept in the directory
- Use Facebook social gaming applications ("Ghostbusting Facebook: Detecting and Characterizing Phantom Profiles in Online Social Gaming Application" - by A. Nazir, S. Raza, C. Chanuch, B. Schipper)
Resources

- Black Hat forums
- Facebook API
- Facebook developers forums

"Prying Data out of Social Networks" (by Joseph Bonneau, Jonathan Anderson and George Danezis).

"Defacing Facebook: A Security Case Study" (by Adrienne Felt, University of Virginia)
Questions?