Thwarting fake accounts by predicting their victims

Yazan Boshmaf
University of British Columbia

Dionysios Logothetis, Georgios Siganos
Telefonica Research

Matei Ripeanu, Konstantin Beznosov
University of British Columbia
Improving fake account detection in OSNs by predicting potential victims

Why are users accounts valuable?

Yazan Boshmaf
University of British Columbia

Dionysios Logothetis, Geogios Siganos
Telefonica Research

Matei Ripeanu, Konstantin Beznosov
University of British Columbia

Submitted to the 18th International Conference on Financial Cryptography and Data Security (FC’14)
User accounts are assets

Average MAU* in Facebook (Millions)**

* Monthly active user (MAU): The basic user metric in Facebook
** Facebook Quarterly Reports, Facebook Investor Relations: http://investor.fb.com

1.2 billion MAU in Q3’13
User accounts generate revenue

Average revenue per Facebook user*

![Average Revenue per User (ARPU)]

Q3'12 Q4'12 Q1'13 Q2'13 Q3'13
Europe

Rest of World

Worldwide

US & Canada

Payments

Advertising

2 billion US dollars in Q3’13

* Facebook Quarterly Reports, Facebook Investor Relations: [http://investor.fb.com](http://investor.fb.com)
How many user accounts are fake?

2 billion US dollars in Q3’13

<table>
<thead>
<tr>
<th></th>
<th>Q3’12</th>
<th>Q4’12</th>
<th>Q1’13</th>
<th>Q2’13</th>
<th>Q3’13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Payments</td>
<td>$1.11</td>
<td>$1.29</td>
<td>$1.15</td>
<td>$1.41</td>
<td>$1.53</td>
</tr>
<tr>
<td>Advertising</td>
<td>$0.25</td>
<td>$0.20</td>
<td>$0.19</td>
<td>$0.19</td>
<td>$0.19</td>
</tr>
</tbody>
</table>

* Facebook Quarterly Reports, Facebook Investor Relations: [http://investor.fb.com](http://investor.fb.com)
Fake accounts are rising

Undesirable* accounts in Facebook (Millions)**

50 million fakes in Q3’13

* Undesirable Facebook accounts include both duplicates and fake accounts (worst case estimates)
** Facebook Quarterly Reports, Facebook Investor Relations: http://investor.fb.com
Fake accounts are rising

Undesirable* accounts in Facebook (Millions)**

<table>
<thead>
<tr>
<th>Quarter</th>
<th>Duplicate</th>
<th>Fake (Benign)</th>
<th>Fake (Malicious)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2'12</td>
<td>47.8</td>
<td>23.9</td>
<td>14.9</td>
</tr>
<tr>
<td>Q3'12</td>
<td>48.3</td>
<td>24.2</td>
<td>15.1</td>
</tr>
<tr>
<td>Q4'12</td>
<td>52.8</td>
<td>13.7</td>
<td>9.5</td>
</tr>
<tr>
<td>Q1'13</td>
<td>55.5</td>
<td>14.4</td>
<td>10.0</td>
</tr>
<tr>
<td>Q2'13</td>
<td>57.8</td>
<td>15.0</td>
<td>10.4</td>
</tr>
<tr>
<td>Q3'13</td>
<td>143.9</td>
<td>25.0</td>
<td>25.0</td>
</tr>
</tbody>
</table>

17.4 thousand fakes per hour on average

* Undesirable Facebook accounts include both duplicates and fake accounts (worst case estimates)
** Facebook Quarterly Reports, Facebook Investor Relations: [http://investor.fb.com](http://investor.fb.com)
Fake accounts are rising

Undesirable* accounts in Facebook (Millions)**

Why are fake accounts harmful?

17.4 thousand fakes removed per hour

* Undesirable Facebook accounts include both duplicates and fake accounts
** Facebook Quarterly Reports, Facebook Investor Relations: [http://investor.fb.com](http://investor.fb.com)
Fake accounts are bad for business

“... If advertisers, developers, or investors do not perceive our user metrics to be accurate representations of our user base, or if we discover material inaccuracies in our user metrics, our reputation may be harmed and advertisers and developers may be less willing to allocate their budgets or resources to Facebook, which could negatively affect our business and financial results...”
Fake accounts are bad for users

OSNs are attractive medium for abusive content*

Free infrastructure to steal data, spread malware & misinform

Bad for users is bad for business

Your Facebook friends may be evil bots
Computer scientists have unleashed hordes of humanlike social bots to infiltrate Facebook -- and they're awfully effective
By Eagle Gamma | InfoWorld

Koobface virus hits Facebook
An e-mail lure and a fake Adobe Flash update request could load a nasty virus on your PC.
by Robert Vamosi | December 4, 2008 4:36 PM PST

'Socialbots' Invade Facebook: Cull 250GB of Private Data
By John P. Mello Jr, PCWorld | Nov 2, 2011 2:20 PM
Fake accounts are bad for users.

What's the role of fake accounts in today's underground economy?
Fake accounts are market enablers

* Keys to many walled gardens*

Fake account are profitable “commodity”

Prices from BuyAccs.com

<table>
<thead>
<tr>
<th>Web Service</th>
<th>Price per Thousand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotmail.com, resale*</td>
<td>$2.00</td>
</tr>
<tr>
<td>Hotmail.com</td>
<td>$4.00</td>
</tr>
<tr>
<td>Yahoo</td>
<td>$6.00</td>
</tr>
<tr>
<td>Twitter</td>
<td>$20.00</td>
</tr>
<tr>
<td>Google (PVA)**</td>
<td>$100.00</td>
</tr>
<tr>
<td>Facebook (PVA)**</td>
<td>$100.00</td>
</tr>
</tbody>
</table>

Already a multi-million dollar business

* Resale indicates account was previously used in another activity
** Phone Verified Accounts: A fake account verified by a text challenge-response using a cell phone
Fake account are profitable “commodity”

How do OSNs fight against fakes?

<table>
<thead>
<tr>
<th>Web Service</th>
<th>Price per Thousand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hotmail.com, resale*</td>
<td>$2.00</td>
</tr>
<tr>
<td>Hotmail.com</td>
<td>$4.00</td>
</tr>
<tr>
<td>Yahoo</td>
<td>$6.00</td>
</tr>
<tr>
<td>Twitter</td>
<td>$20.00</td>
</tr>
<tr>
<td>Google (PVA)**</td>
<td>$100.00</td>
</tr>
<tr>
<td>Facebook (PVA)**</td>
<td>$100.00</td>
</tr>
</tbody>
</table>

* Resale indicates account was previously used in another activity
** PVA indicates a phone verified account; challenge response text to cell phone

Already a multi-million dollar business
Threat model

Attacks can create and control fakes in a botnet-like fashion

Attacking first infiltrate the OSN then mount subsequent attacks

- Account at high Risk
- Victim account
- Account at low risk
- Fake account

Strong tie: ———
Weak tie: ————
Fake-centered security paradigm

Detect fake account by identifying what “fakeness looks like”

- Account at high Risk
- Account at low risk
- Victim account
- Fake account

Strong tie: ________
Weak tie: --------

Defined by anomalies in social content or structure
Feature-based detection

Identifies suspicious accounts using supervised machine learning

Relies on features extracted from real and fake accounts

Which features to use?

- Pictures?
- Friends?
- Posts?
- Interactions?
- Triadic closure?
- Ad clicks?
Which features to use?

- Pictures?
- Friends?
- Interactions?
- Triadic closure?
- Ad clicks?
- Posts?

Is this account fake or real?

* Barreno et al. The security of machine learning. J. on Machine Learning, 2010
Fake accounts $\approx$ real accounts

* Barreno et al. The security of machine learning. J. on Machine Learning, 2010
How to build a ground-truth?

Analysts verify suspicious accounts and update ground-truth

Roadblocks to “quarantine” highly-suspicious accounts

How to build a ground-truth?

Internet crowds verify suspicious accounts to update ground-truth

Roadblocks to “quarantine” highly-suspicious accounts

* Wang et al. Social Turing Tests: Crowdsourcing Sybil Detection. NDSS, 2013
Fake accounts $\approx$ real accounts

- Proactive protection
- Near real-time responses
- Scales to millions of users
- Hard to circumvent
- Accurate detection
- Provably secure

What else can we do?

Graph-based detection

Identifies suspicious accounts using (network) graph analysis

Relies on the structural properties of real and fake accounts

* Boshmaf et al. Graph-based Sybil detection in social and information systems. ASONAM, 2013.
Which structural properties?

Find a (provably) sparse cut between the regions

What about real-world graphs?

A Facebook community of 2,991 user accounts
What about real-world graphs?

Is the community infiltrated?

A Facebook community of 2,991 user accounts
No sparse cut \( \approx \) no fake accounts
But users are easily deceived...

Red denotes fake

Black denotes real

But users are easily deceived...

Users are easily deceived

- Proactive protection
- Near real-time responses
- Scales to millions of users
- Hard to circumvent
- Accurate detection (conditional)
- Provably secure

Can we do better?

Victim-centered security paradigm

Detect fake accounts by first identifying their (potential) victims.

This leads to a more resilient defense mechanism (epidemiology?)
SybilPredict in a nutshell

Predicts victims who (are likely to) have attack edges with fakes in $O(n \log n)$ time

Pros:
- Proactive protection
- Near real-time responses
- Scales to millions of users
- Hard to circumvent

Cons:
- Doesn’t identify fakes
- Introduces usability problems
- Not provably secure
SybilPredict in a nutshell

Embeds predictions into graph to identify suspicious accounts in $O(n \log n + m)$ time

Pros:
- Scales to millions of users
- Hard to circumvent
- Accurate detection
- Provably secure

Cons:
- Retroactive detection
- Batch processed

Uses short random walks biased against identified victims to rank users
SybilPredict in a nutshell

Uses distributed machine learning and graph processing infrastructure

Runs in $O(n \log n + m)$ time end-to-end
We claim SybilPredict is:

- ✔ Proactive protection
- ✔ Near real-time responses
- ✔ Scales to millions of users
- ✔ Hard to circumvent
- ✔ Accurate detection
- ✔ Provably secure
Challenges and research directions

* Boshmaf et al. Key challenges in defending against malicious socialbots. Usenix LEET, 2012
More info?

Fork or clone SybilPredict now: https://grafos.ml

On-going deployment at tuenti

For SybilPredict technical report, please email at boshmaf@ece.ubc.ca
Fork or clone SybilPredict now: https://grafos.ml

Details: Example & prelim results

On-going deployment at tuenti

For latest technical report, please email at boshmaf@ece.ubc.ca
How does it work?

Cut size $= |E_A| = 3$

Non-Sybil region  

Sybil region
Graph-based detection fails now

Cut size $= |E_A| = 10$

Non-Sybil region  Sybil region

Idea: Artificially prune attack edges based of victim prediction
Victim prediction

Trusted

Inspected

Victim

Not-victim

<table>
<thead>
<tr>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Victim prediction

<table>
<thead>
<tr>
<th>Feature vector</th>
<th>Target variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Friends</td>
<td>#Photos</td>
</tr>
<tr>
<td>1</td>
<td>54</td>
</tr>
</tbody>
</table>

```
TP  TN  FP  FN
0   1  0   0
```
Victim prediction

<table>
<thead>
<tr>
<th>#Friends</th>
<th>#Photos</th>
<th>#Posts</th>
<th>Victim?</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>42</td>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>

Feature vector

[Diagram of network with nodes labeled as Trusted, Inspected, Victim, and Not-victim]

<table>
<thead>
<tr>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Victim prediction

<table>
<thead>
<tr>
<th>#Friends</th>
<th>#Photos</th>
<th>#Posts</th>
<th>Victim?</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>512</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

Feature vector

Target variable

- Trusted
- Inspected
- Victim
- Not-victim

TP 1

TN 2

FP 0

FN 0
Victim prediction

<table>
<thead>
<tr>
<th>#Friends</th>
<th>#Photos</th>
<th>#Posts</th>
<th>Victim?</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>476</td>
<td>...</td>
<td>688</td>
</tr>
</tbody>
</table>

Feature vector

Target variable

Trusted

Inspected

Victim

Not-victim

TP | TN | FP | FN
---|----|----|----
2  | 2  | 0  | 0  
Victim prediction

<table>
<thead>
<tr>
<th>#Friends</th>
<th>#Photos</th>
<th>#Posts</th>
<th>Victim?</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>564</td>
<td>…</td>
<td>1</td>
</tr>
</tbody>
</table>

Feature vector

Target variable

- Trusted
- Inspected
- Victim
- Not-victim

<table>
<thead>
<tr>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Victim prediction

<table>
<thead>
<tr>
<th>#Friends</th>
<th>#Photos</th>
<th>#Posts</th>
<th>Victim?</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2066</td>
<td>…</td>
<td>766</td>
</tr>
</tbody>
</table>

Feature vector

Target variable

TP 3  TN 2  FP 0  FN 1

Trusted
Inspected
Victim
Not-victim
Victim prediction

<table>
<thead>
<tr>
<th>Feature vector</th>
<th>Target variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Friends</td>
<td>Victim?</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>#Photos</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
</tr>
<tr>
<td>#Posts</td>
<td></td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
<tr>
<td>58</td>
<td></td>
</tr>
</tbody>
</table>

- Trusted
- Inspected
- Victim
- Not-victim

<table>
<thead>
<tr>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Victim prediction

<table>
<thead>
<tr>
<th>#Friends</th>
<th>#Photos</th>
<th>#Posts</th>
<th>Victim?</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>32</td>
<td>...</td>
<td>58</td>
</tr>
</tbody>
</table>

Feature vector

Target variable

- Trusted
- Inspected
- Victim
- Not-victim

<table>
<thead>
<tr>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Victim prediction

FPR = 8.3%, TPR = 75%
Malicious account detection

Assigns weights to edges based on victim predictions

Cut size = $\text{vol}(E_A) = 1.9$

Penalizes relationships of identified victims (low edge weights)

- Trusted
- Inspected
- Victim
- Not-victim

High = 1
Medium < 1
Low = 0.1
Malicious account detection

Ranks accounts by degree-normalized landing probabilities of a short random walk

Real accounts ≈ similar ranks but malicious accounts ≈ significantly smaller ranks
Malicious account detection

Landing probability

<table>
<thead>
<tr>
<th>Step</th>
<th>Landing probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

```
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```

High = 1
Medium < 1
Low = 0.1
Malicious account detection

<table>
<thead>
<tr>
<th>Step</th>
<th>Legacy probability</th>
<th>Landing probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.4545</td>
<td>0.4545</td>
</tr>
<tr>
<td>3</td>
<td>0.0455</td>
<td>0.0455</td>
</tr>
<tr>
<td>4</td>
<td>0.0455</td>
<td>0.0455</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Malicious account detection

<table>
<thead>
<tr>
<th>Landing probability</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.388</td>
<td>0.1515</td>
<td>0.2273</td>
<td>0.0319</td>
<td>0.0228</td>
<td>0.0046</td>
<td>0.1515</td>
<td>0.0046</td>
<td>~0.98</td>
</tr>
</tbody>
</table>

Step 2

Landing probability

<table>
<thead>
<tr>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0092</td>
<td>0.0046</td>
<td>0.0046</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>~0.02</td>
</tr>
</tbody>
</table>

Early-terminates the random walk after $O(\log n)$ steps
Malicious account detection

Normalized probability

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1764</td>
<td>0.0758</td>
<td>0.0758</td>
<td>0.0319</td>
<td>0.0228</td>
<td>0.0046</td>
<td>0.1515</td>
<td>0.0038</td>
</tr>
</tbody>
</table>

Step

- 

Normalized probability

<table>
<thead>
<tr>
<th></th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0.0028</td>
<td><strong>0.0046</strong></td>
<td><strong>0.0035</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Ranks a node by its degree-normalized landing probability
Malicious account detection

Sorts accounts then estimates a threshold to identify suspicious ones

FPR = 0%
TPR = 88%

Identified as suspicious
Malicious account detection

**Theorem:** Number of fake accounts that rank equal or higher than real accounts is $O(\text{vol}(E_A) \log n)$ where $\text{vol}(E_A) \leq |E_A|$. 

Diagram showing a network with labeled nodes, colored to represent trusted, fake, and real accounts. The nodes are interconnected with lines, and the network includes a range of values from 0.1764 to 0.0038.
Malicious account detection

Theorem: Number of fake accounts that rank equal or higher than real accounts is $O(\text{vol}(E_A) \log n)$ where $\text{vol}(E_A) \leq |A_E|$

How does it perform in practice?
Real-world Sybil activity in Facebook

Data were collected in 2011 January 28 through March 23

A total of 8.8K users received friend requests (32.4% victims)

More mutual friends more likely to accept a request sent by fakes
Real-world Sybil activity in Facebook

Data were collected in 2011 January 28 through March 23

~2K different cities across 127 countries
Real-world Sybil activity in Facebook

Data were collected in 2011 January 28 through March 23

43 different languages
A mean of 5.4 years on Facebook
Real-world Sybil activity in Facebook

Data were collected in 2011 January 28 through March 23

139K nodes, 660K edges, 74 communities, diameter of 9
Predicting victims

Random Forests (RF) is 40% better than random

18 features from public profiles

Random Forests classifier with AUC=0.7
Detecting malicious accounts

Trace-driven simulation on most infiltrated community (3K nodes)

Figure 6: Detecting Sybils: In (e), the distance from the Sybil region is the smallest shortest path from any trusted node to any of the Sybils. In (f), the curves represent rank distributions as computed by SybilPredict using an RF classifier. In particular, the lower curve shows this distribution when the graph contained 2K attack edges (i.e., early detection), while the upper one shows the same distribution when the graph contained 35K attack edges (i.e., late detection).

Few (random) seeds are enough

Near prefect ranking, up to 30% better
Detecting malicious accounts

Trace-driven simulation on most infiltrated community (3K nodes)

Figure 6: Detecting Sybils: In (e), the distance from the Sybil region is the smallest shortest path from any trusted node to any of the Sybils. In (f), the curves represent rank distributions as computed by SybilPredict using an RF classifier. In particular, the lower curve shows this distribution when the graph contained 2K attack edges (i.e., early detection), while the upper one shows the same distribution when the graph contained 35K attack edges (i.e., late detection).

Seeds are sensitive to targeted attacks

Clear cutoff threshold in rank distribution
Detecting malicious accounts

Near linear scalability with exponentially increasing order

Figure 7: Scalability evaluation: In (a), an RF classifier is trained on Mahout. In (b), the social graph is transformed using the predictions evaluated by the RF classifier, after which the nodes are ranked and sorted on Giraph.

Using the same parameters estimated in §6.2–6.3, we ran SybilPredict on the generated workload, ending up with a nearly linear scalability in terms of number of nodes, as shown in Figure 7. Excluding the time required to load the 160M node graph into memory, it takes less than 2 minutes to compute the predictions using the calibrated RF classifier and less than 25 minutes to compute and sort the ranks, making our approach computationally practical even for large OSNs such as Facebook.

26

RF is “embarrassingly parallel”

Ranking is “PageRank scalable”
In conclusion, SybilPredict is:

- Proactive protection
- Near real-time responses
- Scales to millions of users
- Hard to circumvent
- Accurate detection
- Provably secure

(patent disclosure submitted!)