Can You Spot the Fakes?
On the Limitations of User Feedback in Online Social Networks

David Mandell Freeman
Head of Anti-Abuse and Anomaly Detection Relevance

LinkedIn

Perth, Australia
6 April 2017
Fake accounts in social networks

Popular social networks attract bad actors

· scams
· malware
· phishing
· etc.

To carry out abuse, bad guys need fake (or compromised) accounts.

How do we find them?
Reporting fake accounts
Acting on flagging signals

Flagging is a low-precision signal.
  · 35% precision in our LinkedIn data set.

Need to accrue multiple flags before taking action.
  · This takes time.

We could act faster & more accurately if we knew that some flags were more precise than others.

Research question: is there such a thing a “super-flagger”?
How do we test whether “super-flaggers” exist?

If flagging is a real skill, it must be:

**measurable** — possible to distinguish from random guessing

**repeatable** — persistent over repeated sampling
Our contribution

Framework for assessing flagging skill.

Apply framework to LinkedIn data:

- profile report spam
- invitation reject
- invitation accept (signal for real accounts)

Conclusion: skilled flaggers exist but are very rare.

- no noticeable impact on metrics
Prior work

[Zheleva et al. ‘08], [Chen et al. ‘15]: Framework to upweight high-precision reporters in spam classification algorithms, mechanism for reputation to evolve.

- Assumes an initial set of high-precision reporters can be identified.
- Assumes identified reporters will continue to be high-precision.

[Wang et al. ’13], [Cresci et al. ’17]: Crowdsourcing studies.

- “People can identify differences between [fake] and legitimate profiles, but most individual testers are not accurate enough to be reliable.”
- Low accuracy on “social spambots”

[Moore-Clayton ‘08] [Chia-Knapskog ’11]: “wisdom of crowds”

- Frequent reporters have higher accuracy (counter to our findings)
Profile flagging data set

Data: all LinkedIn “fake profile” flags over 6-month period

- 293K flags, 227K reporters, 238K reports
- Anti-Abuse team labeled flagged accounts as real or fake
- 35% overall precision

Precision does not improve with number of flags:
Measurability: Precision

How many flags did the user get right?

\[ P(u) = \frac{\# \text{ correct flags}}{\# \text{ flags}} \]

Problem: insensitive to number of flags

\cdot 1 out of 1 is as good as 50 out of 50

Solution: smoothing

\[ P_s(u) = \frac{\# \text{ correct flags} + \alpha}{\# \text{ flags} + 2\alpha} \]

\cdot find \( \alpha \) by optimizing on a test set
Measurability: Precision

How many flags did the user get right?

\[ P(u) = \frac{\# \text{ correct flags}}{\# \text{ flags}} \]

Problem: insensitive to number of flags

- 1 out of 1 is as good as 50 out of 50

Solution: smoothing

\[ P_s(u) = \frac{\# \text{ correct flags} + \alpha}{\# \text{ flags} + 2\alpha} \]

- find \( \alpha \) by optimizing on a test set

Smoothed Precision of Profile Flaggers
Measurability: Informedness

<table>
<thead>
<tr>
<th>$u$</th>
<th>Report</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Fake</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

precision $= \frac{5}{10} = 0.5$

<table>
<thead>
<tr>
<th>$u'$</th>
<th>Report</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>Fake</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

precision $= \frac{5}{10} = 0.5$
Measurability: Informedness

Precision is insensitive to level of fake account exposure:

<table>
<thead>
<tr>
<th></th>
<th>Report</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Fake</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

\[
\text{precision} = \frac{5}{10} = 0.5
\]

<table>
<thead>
<tr>
<th></th>
<th>Report</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>Fake</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

\[
\text{precision} = \frac{5}{10} = 0.5
\]
Measurability: Informedness

Precision is insensitive to level of fake account exposure:

\[
\begin{array}{c|cc}
 u & \text{Report} & \text{Ignore} \\
\hline
\text{Real} & 5 & 5 \\
\text{Fake} & 5 & 5 \\
\end{array}
\]

\[
\text{precision} = \frac{5}{10} = 0.5
\]

\[
\begin{array}{c|cc}
 u' & \text{Report} & \text{Ignore} \\
\hline
\text{Real} & 5 & 95 \\
\text{Fake} & 5 & 5 \\
\end{array}
\]

\[
\text{precision} = \frac{5}{10} = 0.5
\]

**Informedness:** How much better is the user at flagging fake accounts than real ones?

\[
I(u) = \text{TPR} - \text{FPR} = \frac{\# \text{ flags of fakes}}{\# \text{ fakes seen}} - \frac{\# \text{ flags of reals}}{\# \text{ reals seen}}
\]
Measurability: Informedness

Precision is insensitive to level of fake account exposure:

<table>
<thead>
<tr>
<th></th>
<th>Report</th>
<th>Ignore</th>
<th></th>
<th>Report</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>5</td>
<td>5</td>
<td>Fake</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Precision = \( \frac{5}{10} = 0.5 \)

\[
\text{informedness} = \frac{5}{10} - \frac{5}{10} = 0
\]

\[
I(u) = \text{TPR} - \text{FPR} = \frac{\# \text{ flags of fakes}}{\# \text{ fakes seen}} - \frac{\# \text{ flags of reals}}{\# \text{ reals seen}} = 0.45
\]

Informedness: How much better is the user at flagging fake accounts than real ones?
Informedness of Profile Flaggers

Precision is insensitive to level of fake account exposure:

<table>
<thead>
<tr>
<th></th>
<th>Report</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Fake</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Informedness = \( \frac{5}{10} - \frac{5}{10} = 0 \)

\( I(u) = TPR - FPR = \frac{\# \text{ flags of fakes}}{\# \text{ fakes seen}} - \frac{\# \text{ flags of reals}}{\# \text{ reals seen}} \)
Is it skill or luck?

<table>
<thead>
<tr>
<th></th>
<th>Report</th>
<th>Ignore</th>
<th></th>
<th>Report</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>2</td>
<td>2</td>
<td>Real</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Fake</td>
<td>1</td>
<td>0</td>
<td>Fake</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

informedness = $\frac{1}{1} - \frac{2}{4} = 0.5$

$\text{Fisher's exact test}$ on the 2 x 2 contingency table.

Null hypothesis: user is equally likely to flag real and fake accounts.

$p$-value: probability of finding a matrix “at least as extreme” as $M$. 
Is it skill or luck?

<table>
<thead>
<tr>
<th>(v)</th>
<th>Report</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Fake</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\text{informedness} = \frac{1}{1} - \frac{2}{4} = 0.5
\]

\[
p = 1
\]

<table>
<thead>
<tr>
<th>(v')</th>
<th>Report</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Fake</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\text{informedness} = \frac{10}{10} - \frac{20}{40} = 0.5
\]

\[
p = 0.003
\]

Use a statistical hypothesis test to distinguish the two!

*Fisher’s exact test* on the 2 x 2 contingency table.

Null hypothesis: user is equally likely to flag real and fake accounts.

\(p\)-value: probability of finding a matrix “at least as extreme” as \(M\).
Measurability: Hypothesis Testing

Fisher’s test produces a $p$-value: probability of finding a matrix “at least as extreme” as $M$.

— define “Fisher Score” = $1 - p$-value

Problem: statistically significant flaggers may not be good flaggers

<table>
<thead>
<tr>
<th>$w$</th>
<th>Report</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>20</td>
<td>80</td>
</tr>
<tr>
<td>Fake</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

precision = \( \frac{5}{25} = 0.2 \)

informedness = \( \frac{5}{5} - \frac{20}{100} = 0.3 \)

Fisher score = $1 - 0.05 = 0.95$
Repeatability — Correlation

Are skilled flaggers in data set $A$ the same as skilled flaggers in data set $B$?

*Pearson correlation coefficient*: linear correlation of scores.  
*Spearman correlation coefficient*: Pearson correlation of rank vectors.

<table>
<thead>
<tr>
<th>Flagging Score</th>
<th>Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoothed Precision</td>
<td>0.69</td>
<td>0.66</td>
</tr>
<tr>
<td>Informedness</td>
<td>0.52</td>
<td>0.49</td>
</tr>
<tr>
<td>Fisher Score</td>
<td>0.62</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Problem: independent of score magnitude

<table>
<thead>
<tr>
<th>user</th>
<th>$A$ score</th>
<th>$B$ score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>0.94</td>
<td>0.1</td>
</tr>
<tr>
<td>$b$</td>
<td>0.95</td>
<td>0.2</td>
</tr>
<tr>
<td>$c$</td>
<td>0.96</td>
<td>0.3</td>
</tr>
<tr>
<td>$d$</td>
<td>0.97</td>
<td>0.4</td>
</tr>
<tr>
<td>$e$</td>
<td>0.98</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Perfect correlation!
Probability that user with a good score in data set A also has a good score in data set B?

Define persistence at score $\beta$ to be

$$\pi(\beta) = \frac{\# \text{ users with score } > \beta \text{ in } A \text{ and } B}{\# \text{ users with score } > \beta \text{ in } A \text{ or } B}$$

Persistence on flagging data:
Putting it all together

Compute skill threshold for each measurement based on precision on a held-out test set.

- Threshold is such that error rate is less than half the average.

Define “skilled flagger” to be one who is above the threshold on 2 of 3 metrics, on 2 different data sets

- high smoothed flagging precision
- flags real and fake accounts in different proportion
- difference in behavior in flagging real and fake accounts is statistically significant
Profile flagging — skilled flaggers

5600 skilled flaggers
  · 31% of those who flagged ≥2 times
  · 2.4% of all flaggers
  · 82% cumulative precision

4300 high-precision skilled flaggers
  · 13940 accounts flagged (77/day)
  · 97% cumulative precision
Data set 2: Invitation response

Invitation *reject*: reporting signal on *fake* accounts

Invitation *accept*: reporting signal on *real* accounts

Evaluation:

- 500,000 members from June 2016 receiving \( \geq 2 \) spam and \( \geq 3 \) non-spam invitations
- look at responses within the first 24 hours
- 1.3% were skilled at *rejecting fakes*
- 3.8% were skilled at *accepting reals*
An experiment

Simulation: replace member’s responses to fake accounts with binomial samples distributed like responses to real accounts.

<table>
<thead>
<tr>
<th></th>
<th>Report</th>
<th>Ignore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Fake</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Simulated Fake</td>
<td>~ B(0.25)</td>
<td>2</td>
</tr>
</tbody>
</table>

- Fisher scores are lower for simulated data
- persistence drops to zero much more quickly for simulated data

$\sim B(0.25)$
Conclusions

Motivating question: Are there some social network users who are good at identifying fake accounts?

Answer: yes, but not enough to make acting on the signal worthwhile:

- < 2.4% of profile flaggers
- < 1.3% of members rejecting invitations
- < 3.8% of members accepting invitations (i.e. identifying real accounts)

Further work:

- investigate UI changes to improve flagging ability
- find other features correlated with skill (e.g. geo)
Questions?

dfreeman@linkedin.com