Identifying Spam in the iOS App Store

Speaker: Rishi Chandy
Collaboration with Jay Gu
Outline

• Background
• App Store Datasets
• Observations
• Finding spam
  – supervised, unsupervised
• Future Work
• Conclusion
Background

• 2008 to now: 500k+ apps on Apple app store

• iPhone, iPad restricted to official apps only

• Popular app = millions of dollars (or private data)
The Wrong Way: Path Uploads iOS Users’ Address Books Without Permission

CHRIS VELAZCO

Tuesday, February 7th, 2012

27 Comments
The Problem: Fake Reviews

• Motivation: Fun & Profit
  – attack competitors
  – promote own apps

• Profit
  – Money
  – Private data (address book, location, ...)
Hole 9 crash every time (v3.02) ★★★★★
by Blake1775 - Version 3.0.3 - Dec 4, 2011

I want my money back! Do not buy this app.

2 out of 2 customers found this review helpful

Course updates ★★★★★
by golfer1under - Version 3.0.3 - Nov 10, 2011

Fast and accurate

2 out of 4 customers found this review helpful

Not Worth a Dime ★★★★★
by OldGuy52 - Version 3.0.3 - Jan 29, 2012

Crashes constantly, data lost, no recovery. No response from "Customer Service". I want my money back.

1 out of 1 customers found this review helpful
Previous Work

• None for app store specifically

• Other contexts
App Store Review Graph

Users → Reviews → Apps → Creators

...
Outline

• Background

App Store Datasets

• Observations

• Finding spam
  – supervised, unsupervised

• Future Work

• Conclusion
Datasets

• User-Review-App-Developer relationships

• App metadata
  – name, release date, price, ...

• Reviews
  – user id, timestamp, rating, “helpfulness”
Datasets

- Top Apps (TA), Entertainment & Lifestyle (EL)

<table>
<thead>
<tr>
<th>(approx)</th>
<th>TA</th>
<th>E&amp;L</th>
<th>Labeled E&amp;L</th>
</tr>
</thead>
<tbody>
<tr>
<td># Apps</td>
<td>700</td>
<td>2k</td>
<td>114</td>
</tr>
<tr>
<td># Reviews</td>
<td>6 million</td>
<td>37k</td>
<td>33k</td>
</tr>
<tr>
<td># Users</td>
<td>4 million</td>
<td>37k</td>
<td>33k</td>
</tr>
<tr>
<td># Devs.</td>
<td>400</td>
<td>2k</td>
<td>114</td>
</tr>
</tbody>
</table>
Observations

Log CCDF

Avg Rating per User

1 2 3 4 5

0.75
Outline

• Background

• App Store Datasets

• Observations

Finding spam

  – supervised, unsupervised

• Future Work

• Conclusion
Finding Spam: Supervised

• We manually labeled 114 apps

• Decision Tree

• Simple graphical model
Simple Graphical Model
Simple Graphical Model

<table>
<thead>
<tr>
<th>Node</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>f_u</td>
<td>avg rating, num reviews</td>
</tr>
<tr>
<td>f_a</td>
<td>avg rating, num reviews</td>
</tr>
<tr>
<td>f_r</td>
<td>l(score)</td>
</tr>
<tr>
<td>f_d</td>
<td>avg rating, num apps</td>
</tr>
</tbody>
</table>
Finding Spam: Supervised

• We manually labeled 114 apps

• Decision Tree
  – 41% error

• Simple graphical model
  – 27% error
Finding Spam: Unsupervised

- Unlabeled!

- Cluster using our simple model
Clustering Reviews

![Graph showing the relationship between Pr(Ir=0) and Pr(Ir=1).]
Clustering Apps
Clustering Developers
Outline

• Background
• App Store Datasets
• Observations
• Finding spam
  – supervised, unsupervised

Future Work

• Conclusion
Future Work

• Larger dataset

• LDA on review text

• Temporal analysis
Conclusion

• App store spam: important problem

• Baseline methods have decent performance

• Huge opportunities for future work