Click Fraud Detection using Practical Memetics

Defcon 15, Aug 3-5, 2007
Broward Horne

http://www.realmeme.com
Overview

• Advertising Click Fraud
• Botnets (Agents Of Click Fraud)
• S-Curve
• Original Meme Theory (Dawkins)
• Empirical Meme Theory (Me! Ho!)
• Expanded Meme Mining Model
• MySpace Example
• IAFF.com Example
• Meme Seepage Theory
• Botnet Proof
• Gaming Conjectures
• Extrapolations
Click Fraud - Definition

From Wikipedia -

“When a person, script or computer program imitates a legitimate user to generate a “charge per click” of an advertised product”

Is Click Fraud relatively unknown? Poll
Google Adsense Model

- www.google.com/adsense
- Click fraud risk
Botnets

- From Wikipedia-

  “a collection of compromised computers ("zombies") running... under a common command-and-control infrastructure”
S-Curve (Physical Model)
S-Curve (Math Model)
S-Curve Rate-Of-Change
S-Curve Strategy
Dawkins Meme Theory

• Dawkins coined the term “meme” in 1976
• An idea like “I want a tattoo”
• Wikipedia definition -

“a unit of cultural information that propagates from one mind to another as a theoretical unit of cultural evolution”
Ideosphere

- The sum of all memes in circulation
- The “Global Human Consciousness”
- The Internet has a subset of the Ideosphere
Empirical Meme Theory (Me!)

- Original Meme Theory
- Keywords As Proxy For A Meme
- Electron Flow \( (E = I \times R) \) (Networks)
- S-Curve
- Meme Miner (Dejanews.com)
- Google Trends tool
- Blogpulse.com
Example: Delphi
Example: Easter Bunny

Dejanews MemeGraph For: 'easter bunny'

Meme Miner 1.1
http://wwwrealmeme.com/miner

Frequency Count


easter bunny easter bunny moving average
Example: Sex & Terrorism

Dejanews MemeGraph For: 'terrorism vs. sex'

Meme Miner 1.1

http://www.realmeme.com

Ideosphere deallocates bandwidth for sex, allocates it to terrorism instead

Dejanews Goes Bad Here

Sex

Terrorism

Frequency Count

Jan 1995
Jul 1995
Jan 1996
Jul 1996
Jan 1997
Jul 1997
Jan 1998
Jul 1998
Jan 1999
Jul 1999
Jan 2000
Jul 2000
Jan 2001
Jul 2001
Jan 2002
Jul 2002
Jan 2003
Jul 2003
Jan 2004
Jul 2004
Jan 2005
Jul 2005
Jan 2006

[Diagram showing frequency count over time with two lines labeled 'terrorism' and 'sex', and a note: 'Ideosphere deallocates bandwidth for sex, allocates it to terrorism instead.']
Meme Assumptions

• Memes propagate as an S-Curve
• Memes propagate to most sites but at different amplitudes and latencies
Meme Miner Inadequacies

- Single source
- Dejanews.com indexing revised by Google
- Older technology losing favor
Expanded Mining Model

First tested on “MySpace” meme…
MySpace Meme (Dejanews)

Dejanews MemeGraph For: 'myspace'

Meme Miner 1.1
http://www.realmeme.com

An unsustainable rate of change.
This baby is topping out
MySpace Meme (Alexa)

Inflection Point
MySpace Meme (Google)
The New Model Worked…

For the “MySpace” meme, so I tested it against another case, a new site with a high growth rate…

“IAmFacingForeclosure.com” (IAFF.com)

But the results were different…
IAFF Meme (Alexa)

Failure to break previous high is a likely confirmation of downward trend

"national broadcast event"
IAFF Meme (Dejanews)

Dejanews MemeGraph For: 'iamFacingForeclosure.com'

Meme Miner 1.1

http://www.realmeme.com

Almost a zero count
IAFF Meme (Google)
Meme Seepage Theory

Remember our Meme assumptions? Memes propagate as an S-curve and across most sites but with varying amplitudes and latencies.

If traffic increases to a primary site, then traffic to linked sites should increase proportionally (more or less)

If traffic increases to a primary site, then traffic to reference sites like Google should increase proportionally… and other sites in the top 10 result list should experience a lesser but measurable increase in traffic…
So I Experimented..

I posted a link directly to IAFF.com, to tap off a slice of IAFF.com’s traffic via meme seepage. Theoretically, a doubling of IAFF traffic should produce an equal ratio of redirected traffic to my own site, RealMeme.com

But the results were wrong again…
And I Experimented Again…

Alexa showed IAFF.com with a 25% increase in traffic but my site experienced no concurrent increase.

I was surprised and I posted the results to IAFF.com.

And a few days later, I did experience an anomalous increase in traffic but it didn’t match IAFF.com’s Alexa traffic delta. Here’s what hit my site…
Botnet Proof

The following page hits are logs from my website. It’s clear that they were artificially produced…

- Too many simultaneous operating systems per IP.
- The traffic is too dense and changeover too abrupt.
- Too many 2-page hits (my traffic is 90% 1-page)
- The page hits don't follow a human click flow.
- All hits are bookmarks, no blog entry points
- Most bookmarks are older
Botnet Logs (Same IP)

With six different operating systems hitting my site simultaneously

And hitting again
Botnet Logs (Density)

<table>
<thead>
<tr>
<th>IP Address</th>
<th>dows Marketplace</th>
<th>Windows Media</th>
<th>Windows</th>
</tr>
</thead>
<tbody>
<tr>
<td>75.108.41.55</td>
<td>2</td>
<td>2</td>
<td>36.28 KB</td>
</tr>
<tr>
<td>70.232.141.251</td>
<td>4</td>
<td>4</td>
<td>73.43 KB</td>
</tr>
<tr>
<td>24.3.2.67</td>
<td>1</td>
<td>1</td>
<td>18.19 KB</td>
</tr>
<tr>
<td>68.81.10.157</td>
<td>1</td>
<td>1</td>
<td>77.86 KB</td>
</tr>
<tr>
<td>70.248.168</td>
<td>2</td>
<td>2</td>
<td>37.67 KB</td>
</tr>
<tr>
<td>24.215.125</td>
<td>1</td>
<td>1</td>
<td>17.93 KB</td>
</tr>
<tr>
<td>67.101.170.151</td>
<td>2</td>
<td>2</td>
<td>39.09 KB</td>
</tr>
<tr>
<td>66.41.129.130</td>
<td>2</td>
<td>2</td>
<td>35.42 KB</td>
</tr>
<tr>
<td>65.29.215.15</td>
<td>4</td>
<td>4</td>
<td>77.86 KB</td>
</tr>
<tr>
<td>168.103.112.71</td>
<td>2</td>
<td>2</td>
<td>36.29 KB</td>
</tr>
<tr>
<td>24.61.4.214</td>
<td>2</td>
<td>2</td>
<td>63.45 KB</td>
</tr>
<tr>
<td>68.98.199.197</td>
<td>2</td>
<td>2</td>
<td>39.04 KB</td>
</tr>
<tr>
<td>74.13.55.126</td>
<td>4</td>
<td>4</td>
<td>74.84 KB</td>
</tr>
<tr>
<td>65.26.19.37</td>
<td>2</td>
<td>2</td>
<td>39.76 KB</td>
</tr>
<tr>
<td>24.203.144.9</td>
<td>2</td>
<td>2</td>
<td>37.13 KB</td>
</tr>
<tr>
<td>71.233.231.234</td>
<td>2</td>
<td>2</td>
<td>38.86 KB</td>
</tr>
<tr>
<td>69.182.232.138</td>
<td>2</td>
<td>2</td>
<td>36.47 KB</td>
</tr>
<tr>
<td>75.21.160.168</td>
<td>2</td>
<td>2</td>
<td>38.65 KB</td>
</tr>
<tr>
<td>24.29.135.227</td>
<td>2</td>
<td>2</td>
<td>37.79 KB</td>
</tr>
<tr>
<td>67.191.202.24</td>
<td>1</td>
<td>1</td>
<td>17.92 KB</td>
</tr>
<tr>
<td>68.255.78.224</td>
<td>2</td>
<td>2</td>
<td>38.87 KB</td>
</tr>
<tr>
<td>70.59.211.122</td>
<td>2</td>
<td>2</td>
<td>36.86 KB</td>
</tr>
<tr>
<td>66.27.14.212</td>
<td>1</td>
<td>1</td>
<td>18.39 KB</td>
</tr>
<tr>
<td>68.252.58.24</td>
<td>1</td>
<td>1</td>
<td>17.66 KB</td>
</tr>
</tbody>
</table>

- Traffic is too dense and change too abrupt
- Too many 2-page hits, 90% of my traffic is 1 page, 5% is more than 2.
- Page sizes are too close but not identical
### Botnet Logs (Pages)

#### Statistics for realmeme.com (2006-12) - Mozilla Firefox

- **Summary**
  - Monthly history
  - Days of month
- **Navigation**
  - Visits duration
  - File type
  - Viewed
  - Full list
  - Entry
  - Exit
  - Operating Systems
    - Versions
    - Unknown
  - Browsers
    - Versions
    - Unknown
- **Referers**
  - Origin
    - Referring search engines
    - Referring sites
  - Search

#### Last visit

<table>
<thead>
<tr>
<th>IP Address</th>
<th>Pages</th>
<th>Hits</th>
<th>Bandwidth</th>
<th>Last visit</th>
</tr>
</thead>
<tbody>
<tr>
<td>219.27.152.55</td>
<td>1</td>
<td>1</td>
<td>7.77 KB</td>
<td>16 Dec 2006 - 05:12</td>
</tr>
<tr>
<td>211.107.250.35</td>
<td>5</td>
<td>5</td>
<td>46.55 KB</td>
<td>16 Dec 2006 - 04:50</td>
</tr>
<tr>
<td>222.101.209.71</td>
<td>2</td>
<td>2</td>
<td>15.11 KB</td>
<td>16 Dec 2006 - 04:50</td>
</tr>
<tr>
<td>219.121.119.43</td>
<td>4</td>
<td>4</td>
<td>114.79 KB</td>
<td>16 Dec 2006 - 04:50</td>
</tr>
<tr>
<td>66.98.186.40</td>
<td>1</td>
<td>1</td>
<td>12.20 KB</td>
<td>16 Dec 2006 - 04:50</td>
</tr>
<tr>
<td>221.241.160.111</td>
<td>4</td>
<td>4</td>
<td>111.35 KB</td>
<td>16 Dec 2006 - 04:49</td>
</tr>
<tr>
<td>210.113.30.208</td>
<td>1</td>
<td>1</td>
<td>7.73 KB</td>
<td>16 Dec 2006 - 04:47</td>
</tr>
<tr>
<td>72.8.86.169</td>
<td>1</td>
<td>1</td>
<td>16.45 KB</td>
<td>16 Dec 2006 - 04:47</td>
</tr>
<tr>
<td>68.81.240.5</td>
<td>4</td>
<td>4</td>
<td>42.00 KB</td>
<td>16 Dec 2006 - 04:47</td>
</tr>
<tr>
<td>194.247.241.213</td>
<td>1</td>
<td>8</td>
<td>102.11 KB</td>
<td>16 Dec 2006 - 04:44</td>
</tr>
<tr>
<td>220.37.184.40</td>
<td>7</td>
<td>7</td>
<td>205.51 KB</td>
<td>16 Dec 2006 - 04:31</td>
</tr>
<tr>
<td>81.29.194.189</td>
<td>2</td>
<td>2</td>
<td>29.57 KB</td>
<td>16 Dec 2006 - 04:11</td>
</tr>
<tr>
<td>66.174.92.162</td>
<td>341</td>
<td>990</td>
<td>15.05 MB</td>
<td>16 Dec 2006 - 04:07</td>
</tr>
</tbody>
</table>
Botnet Logs (Entry Points)

Only bookmarks, not entry points into blog
Mostly older bookmarks, usually with only one reference
Botnet Epiphany

Okay, I’m not the smartest guy in the world but I eventually figured out that this new traffic was generated by bots.

But why?

So I tried another experiment…
Secondary Seepage Failure

I have a confession. My website was designed specifically for Google rankings and it’s been surprisingly successful (my Defcon 16 presentation! Ho!) So I decided…

to induce a secondary seepage from IAFF.com to my site via Google…
Binding

I bound my site to IAFF.com by posting an IAFF.com analysis which was indexed by Google. At one point, I was the #7 Google result for “IAmFacingForeclosure.com”

So now I’m getting a slice of traffic directly from IAFF.com AND from Google’s search results for IAFF.com
Binding Results

Once again, I saw major anomalies between IAFF.com’s claimed traffic and the induced seepage to my site during “a major television event”.

Can I prove fraud? No.

But I don’t need to. I’m not Google or a Google advertiser.
My Theory

The botnets are mimicking meme seepage by generating traffic to linked secondary sites.

At first, I couldn’t figure out why. But as I worked out a methodology to expose botnet manipulation, I realized that LACK OF SEEPAGE is a major red flag. After all, that’s how I found these anomalies to begin with.
Gaming Google

Botnets click embedded Adsense links on IAFF.com and generate the illusion of high traffic to cash in on "Pay-Per-Click" revenue.
Gaming Alexa

A certain percentage of hijacked Bot systems have Alexa toolbars. These systems generate an illusion of increased IAFF.com traffic to Alexa.com
Gaming My Miner Model

Gaming The 3-Part Model

Master Bot

Bot

Bot

Bot [Alexa]

Bot

IAFF.com

Adsense Link

Alexa.com

Google / Trends

Adsense Advertiser
## Meme Troubleshooting Table

<table>
<thead>
<tr>
<th>Communication Avenue</th>
<th>Measurement Avenue</th>
<th>Reference Avenue</th>
<th>Assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dejanews</strong></td>
<td><strong>Alexa.com</strong></td>
<td><strong>Google</strong></td>
<td></td>
</tr>
<tr>
<td>Meme Appears</td>
<td>Meme Appears</td>
<td>Meme Appears</td>
<td>Normal Meme Behavior, i.e. propagation looks like an S-curve</td>
</tr>
<tr>
<td>Meme Appears</td>
<td>Meme Appears</td>
<td>Nothing</td>
<td>An anti-meme, i.e. &quot;IAmFacingForeclosure.com&quot; a train wreck that no one really wants to see</td>
</tr>
<tr>
<td>Meme Appears</td>
<td>Nothing</td>
<td>Meme Appears</td>
<td>Reference node has been hacked, i.e. &quot;gaming Google&quot;</td>
</tr>
<tr>
<td>Meme Appears</td>
<td>Nothing</td>
<td>Nothing</td>
<td>Overdriven Meme - communication model is being overflowed in a brute-force attempt at meme propagation</td>
</tr>
</tbody>
</table>
Humans versus Bots

Needed: a pervasive, immutable quality which is detectable in humans but which bots can never duplicate.

What is it?

Humans actually buy advertised products.
Conclusion

• Click fraud is more pervasive than reported
• It can be detected with memetic analysis…
• Botnets are a serious problem
• It will eventually become almost impossible to detect sophisticated bots.
• What will Google do?