The Anatomy of Comment Spam

Shelly Hershkovitz, Sr. Security Research Engineer, Imperva
Agenda

- Comment Spam - What & Why?
- Comment Spam Attacks
- Data Analysis
- Mitigation Techniques
- Case Studies
- Conclusion
- Q&A
Shelly Hershkovitz, Sr. Security Research Engineer, Imperva

- Leads the efforts to capture and analyze hacking activities
  - Authored several Hacker Intelligence Initiative (HII) Reports
- Experienced in machine learning and computer vision
- Holds BA in Computer Science & M.Sc degree in Bio-Medical Engineering
Comment Spam - What & Why?

- **What?**
  - Wikipedia: "Comment spam is a term used to refer to a broad category of spam bot postings which abuse web-based forms to post unsolicited advertisements as comments on forums, blogs, wikis and online guest books."

- **Why?**
  - Search engine optimization
  - Advertisements
  - Malware distribution
  - Click fraud
Search Engine Optimization

- MyWebSite.com
- OtherWebSite.com
- OtherBlog.com
- OtherWebSite.com
- OtherNewsWebSite.com
Comment Spam Attack

- Target Acquisition
- Comment Generation
- Posting
- Verification
Comment Spam in Practice

- Success relies on large scales
- Automated tools are used
- Inputs
  - The site to be promoted
  - Relevant keywords
Target Acquisition

- **URL Harvesting**
  - Locate relevant websites
  - Locate suitable URLs for commenting
- **An alternative – buy ‘Quality URLs’ lists**
  - A typical price is $40 for ~13,000 URLs
Selecting the Targets

- **Relevance**: Relevance to the promoted site
- **Quality**: The URL’s own search engine ranking
- **Difficulty**: The difficulty of posting comments (Captcha)
- **Policy**: The site’s policy regarding search engine (follow/nofollow attribute)
Target Acquisition in Action
Comment Generation

- Verbal comments attached to the promoted site
  - Input keywords

Don't forget to leave your comments!
Comment Generation in Action

'Spintax' phrase:

{Reading|Studying} {books|papers} can be {interesting|enriching}.

Optional comments:
- Reading books can be interesting.
- Studying papers can be enriching.
Posting

- Post comments on many URLs
- Authentication, CAPTCHA, or user details handling
Posting in Action
Verification

- Collect feedback whether or not the comments were posted
Verification in Action

![ScrapeBox DoFollow/NoFollow Check](image)

<table>
<thead>
<tr>
<th>URL</th>
<th>DoFollow</th>
<th>NoFollow</th>
<th>Assumption</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://cooking-solutions.com/blog/k-rock-comes-to-cooking-class/cooking-class-003/">http://cooking-solutions.com/blog/k-rock-comes-to-cooking-class/cooking-class-003/</a></td>
<td>1</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://www.thefanboyseo.com/comics/spider-land-spoilers-and-teasers-parker-reveals">http://www.thefanboyseo.com/comics/spider-land-spoilers-and-teasers-parker-reveals</a></td>
<td>1</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://seabung.com/ways-to-improve-sales-through-your-website/">http://seabung.com/ways-to-improve-sales-through-your-website/</a></td>
<td>1</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://blog.lanephotographyonline.com/2011/06/02/paduah-engagement-photograp">http://blog.lanephotographyonline.com/2011/06/02/paduah-engagement-photograp</a></td>
<td>2</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://learnmtonfitness.com/blog/4-minute-tabata-workout-by-nick-tumminello">http://learnmtonfitness.com/blog/4-minute-tabata-workout-by-nick-tumminello</a></td>
<td>5</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://blog.toiletpaperworld.com/ts-get-to-know-your-customers-day/">http://blog.toiletpaperworld.com/ts-get-to-know-your-customers-day/</a></td>
<td>6</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://burymerthisdress.com/blog/2006/02/hectic/">http://burymerthisdress.com/blog/2006/02/hectic/</a></td>
<td>1</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://burymerthisdress.com/blog/2005/09/allshook-up/">http://burymerthisdress.com/blog/2005/09/allshook-up/</a></td>
<td>2</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://blog.toiletpaperworld.com/extra-5-off-for-labor-day/">http://blog.toiletpaperworld.com/extra-5-off-for-labor-day/</a></td>
<td>6</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://blog.toiletpaperworld.com/vermbbrand-changing-to-boardwalk-brand/">http://blog.toiletpaperworld.com/vermbbrand-changing-to-boardwalk-brand/</a></td>
<td>1</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://www.beachweddingsbydeb.com/wedding-blog/?p=350">http://www.beachweddingsbydeb.com/wedding-blog/?p=350</a></td>
<td>1</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://cooking-solutions.com/blog/k-rock-comes-to-cooking-class/cooking-class-001/">http://cooking-solutions.com/blog/k-rock-comes-to-cooking-class/cooking-class-001/</a></td>
<td>4</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://burymerthisdress.com/blog/2011/04/are-we-fashion-forward/">http://burymerthisdress.com/blog/2011/04/are-we-fashion-forward/</a></td>
<td>1</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://burymerthisdress.com/blog/2005/09/herewe-go-again/">http://burymerthisdress.com/blog/2005/09/herewe-go-again/</a></td>
<td>3</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://blog.toiletpaperworld.com/23-march-madness-winner-boardwalk-2-play/">http://blog.toiletpaperworld.com/23-march-madness-winner-boardwalk-2-play/</a></td>
<td>4</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://burymerthisdress.com/blog/2005/08/dont-feel-like-going-ugh/">http://burymerthisdress.com/blog/2005/08/dont-feel-like-going-ugh/</a></td>
<td>1</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://lnnsaytanyphoto.com/blog/2011/07/11/wedding-photography-asbury-park-sne">http://lnnsaytanyphoto.com/blog/2011/07/11/wedding-photography-asbury-park-sne</a></td>
<td>1</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://www.spicybugs.com/2011/06/20/good-food-good-friends/">http://www.spicybugs.com/2011/06/20/good-food-good-friends/</a></td>
<td>3</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://redneckbarandgrill.com/our-amazing-videos/redneck-explosions/">http://redneckbarandgrill.com/our-amazing-videos/redneck-explosions/</a></td>
<td>4</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://blog.whitewiskersphotography.com/2010/02/22/south-shore-pet-photography">http://blog.whitewiskersphotography.com/2010/02/22/south-shore-pet-photography</a></td>
<td>1</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
<tr>
<td><a href="http://blog.toiletpaperworld.com/creative-toilet-paper-storage-becomes-toilet-paper-art">http://blog.toiletpaperworld.com/creative-toilet-paper-storage-becomes-toilet-paper-art</a></td>
<td>1</td>
<td>0</td>
<td>DoFollow</td>
<td>Completed, WP</td>
</tr>
</tbody>
</table>

- **Detected DoFollow**: 4 (52%)
- **Detected NoFollow**: 3 (48%)
- **Requests Done**: 100%

- **KB Downloaded**: 19476KB
- **Average URL’s/Second**: 1
- **Connections**: 4
- **Start**: 0
- **Restart Failed**: 0
- **Abort**: 0
- **Export**: 0
- **Exit**: 0
- **Proxies**: 10
- **Connections**: 0

© 2014 Imperva, Inc. All rights reserved.
Comment Spam in Action
Data Analysis

- 17% of the attackers generated 58% of comment spam traffic
Data Analysis

- 80% of comment spam traffic is generated by 28% of attackers
Mitigation Techniques

- Content inspection
- Source reputation
- Anti-automation
- Demotivation
- Manual inspection
Mitigation Techniques: Content Inspection

- Inspecting the content of the posted comments
- Rule based
  - Large number of links
  - Logical sentences not related to the subject
- Akismet
Mitigation Techniques: Source Reputation

- Based on the reputation of the poster
- Online repositories based on crowdsourcing
Mitigation Techniques: Anti-Automation

- Anti-automation tools
  - CAPTCHA
  - Check-box for posting the comment
  - Client type classification
Mitigation Techniques: Demotivation

- Make comment spam useless
- Follow/nofollow value of the rel attribute of an HTML anchor <A>
  - Specifies whether a link should be followed by search engines

```html
<a rel="nofollow" href="http://www.bestCars.com">Best cars!</a>
```

- Penguin update for Google search engine algorithms
Mitigation Techniques: Manual Inspection

- Effective but not scalable
- Effective against manual comment spam
Case Studies

- Attack Target: Specific Victim
- Attack Source: Specific Attacking IP
- Google App Engine
Specific Victim

- A non-profit organization
- A single host with many URLs
- Our theory associates popular phrases within the URL address and page content, to the attack rate

![Chart showing number of attacks per URL]
Specific Victim

- 52% of source IPs produce 80% of the traffic
Specific Attacking IP

- Comment spam posting from a specific IP
- Rapid response (IP reputation feed) would have significantly reduce the impact of the attack

![Graph showing the number of attacks over days from 8/31/17 to 9/16/17]
Specific Attacking IP

- Five target websites were attacked from this source
- Most had suffered a relative high amount of comment spam attacks

![Percentage of Traffic per Target](chart.png)
Specific Attacking IP

- Hyperlinks in a single request are for different websites
- Consecutive requests have similar hyperlinks
- Using different URLs for the same website avoids bad reputation
Case Studies: Google App Engine

- Google App Engine can be used to spread comment spam through proxy services

This technique can be used to bypass IP based mitigations
Conclusion

- Comment spam is a prosperous industry
  - Many tools and services are available for comment spam generation and distribution

- Identifying the attacker as a comment spammer early on and blocking its requests prevents most of the malicious activity
  - Reputation based controls are effective (IP / source application)

- Reputation based controls must be combined with some content based controls to avoid false positives

- Anti-automation and bot-detection controls can reduce the likelihood of an application becoming a target
Webinar Materials

Join Imperva LinkedIn Group, Imperva Data Security Direct, for…

- Post-Webinar Discussions
- Answers to Attendee Questions
- Webinar Recording Link
- Join Group

Join Imperva LinkedIn Group, Imperva Data Security Direct, for…