deSEO: Combating Search-Result Poisoning

Yu Jin @MSCS USF
Your Google is not SAFE!

SEO Poisoning - A new way to spread malware!

Security Software - Free Software Downloads and Software Reviews...
download security center. Your source for antispyware and security downloads ... Software maker will release its Microsoft Security Essentials "in the coming ... Vista and Windows Server 2008, but not the final version of Windows 7. ...
download.cnet.com/windows/security-software/- Similar

Linda Chong's Blog: Download a free copy of Windows Security ...
30 Sep 2009 ... Microsoft Security Essentials is officially released in 8 languages and 19 countries around the world. You can download Microsoft Security ... blogs.msdn.com/.../download-a-free-copy-of-windows-security-essentials-to-protect-your-home-pc-and-laptop-today.aspx - 9 hours ago - Similar

Microsoft Security Essentials Download
29 Sep 2009 ... Free Microsoft Security Essentials available for download. Microsoft has good reason to ensure Windows PCs are secure and malware-free ... www.security-essentials-download - 18 hours ago - Similar
Why choose SE?

22.4% of Google searches in the top 100 results
> 50% for very-popular key phrase search in first page

Why attackers are attracted by search engines?

- Low cost
- Legitimate appearance
  - People trust Google, Bing, Yahoo, etc
Background of SE

- **Page Rank**
  - rank the web pages in search index
  - depends on number of incoming links and ranks of links on pages

- **Features on pages**
  - disclose, to prevent spammers
  - some estimated features:
    - over 200 features
    - ex. words in the title, URL, content of the page.

- **Search Engine Optimization (SEO)**
  - optimize Web pages so that they are ranked higher by search engines.
  - **white-hat**
    - created by the end user primarily
    - sitemap, appropriate headings and subheadings
    - follow guidelines recommended by SE
  - **black-hat**
    - EVIL!
Black-Hat SEO

- Gaming the rankings, do not follow guidelines
  - Keyword stuffing
    - filling the page with lots of irrelevant keywords
  - Cloaking
    - providing different content to crawlers and users
  - Redirects
  - participating in link farms

- Detect Black-Hat SEO pages
  - Content of pages
  - Presence of cloaking
  - link structures leading to the pages

- SEO attacks studied in this paper (Trojan.FakeAV case)
  - Mixed behavior, hard to detect, less training data
    - Servers are originally legitimate
    - Main websites operate normally even after compromised
Overview of SEO Attack

- How a victim falls into trap of SEO poisoning
- Three major players:
  - Compromised Web Server, Redirection Server, Exploit Server
Compromised Server

Attacker

query for oCommerce

Web Server

Non-malicious page for SE bots (cloaking)

Fake Page images/page.php?page=<keys>

SEO Page

Upload Files:
- php scripts
- graphic shell
- ex. images/

Site A

Site B

Site C
SEO Page

Obfuscated PHP → Eval() → Un-obfuscated Code

Is SE? → Yes →

User-Agent String

Logs:
- time,
- user-agent,
- url, etc

Generate Links & Content:
- 40 links to same server from key.txt
- 5 links to other domain
- Relevant content to key phrase from Google top 100 results (url, snippets)
Redirecting & Exploit Server

Compromised Domain

Compromised Domain

Exploit Server
(Fake AV stored)

Cache URL

NailCash

Redirect URL:
feed2.fancyskirt.com/<paras>
cmd=getTdsUrl
productID=3 (FakeAV)
Observations

Figure 3: The number of other compromised sites each site links to. The degree distribution indicates a dense linking structure.

Figure 4: The number of sites compromised by the attackers each day over a period of three months.

Figure 5: The interval between a site getting compromised and the SEO page getting crawled by a search engine.

Figure 6: The frequency with which each keyphrase occurs across the compromised sites.
Observations

- **Link Structure**: Num. of out links less, more likely be compromised
- **Time line**: most at initial phase of attack
  - duration before first SE crawl: half < 4hrs; over 85% < 1day
    - SE crawlers are aggressive
    - Attacker submit to SE
- **Distribution of key phrase**: High rank has high appearance rate
- **Traffic from victims**: over 5000 compromised sites, over 40 million SEO pages
  - by monitoring logs of Redirect Servers (num. of visits of Fake AV)

![Graph](image)

*Figure 7: The arrival of requests at the redirection server.*
Why SEO Attacks succeed?

- 3 key observations:
  - Generation of pages with relevant content
  - Targeting multiple popular search keywords to increase coverage
  - Creating dense link structures to boost pagerank

- deSEO focus on:
  - Websites change behavior
    - many new links are added after compromised, usually with different URL structures from the old URLs
  - Harmful URLs have similar structures
    - looking for newly created pages that share the same structure on different domains
    - help to identify group attacks
History-based Detection

- Study and compare URLs of server with snapshot history

  - Do website have url: /images/news.php?page= before?
Clustering of Suspicious Domains

- Three Lexical features from URLs:
  - String features:
    - separator between keywords, argument name, filename, subdirectory name before the keywords
  - Numerical features:
    - number of arguments in the URL, length of arguments, length of filename, length of keywords
  - Bag of words:
    - keywords

- abcd.blogspot.com v.s. blogspot.com
  - Sub-domains are preferred

- K-means++ method
  - Neither weight and threshold selection are sensitive
Group Analysis

- SEO links in one campaign share a similar page structure (not just the URL structure)
- Measure similarity of web pages
  - Compare parsed HTML structure tree
- AutoRE - signature generator system

Figure 9: An example legitimate group that has diverse distribution of number of URLs in each Web page.

Figure 10: An example malicious group that has a similar number of URLs in each Web page.
Results

- **Data sets (Bing.com):**
  - June, 2010 - Historical snapshots
  - Sep, 2010
  - Jan, 2011

- **Trendy keywords:**
  - Google Trends, May 28th, 2010 ~ Feb, 3rd, 2011 (20/day)

- **History-based Detection:**
  - over 100 billion URLs at beginning
  - filter by Alex Top 10000
  - 2/3 move to next step

<table>
<thead>
<tr>
<th>Month</th>
<th>With trendy keyword</th>
<th>With new structure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Domains</td>
<td>URLs</td>
</tr>
<tr>
<td>Sept 10</td>
<td>428,430</td>
<td>1,481,766</td>
</tr>
<tr>
<td>Jan 11</td>
<td>512,617</td>
<td>3,255,140</td>
</tr>
</tbody>
</table>

Table 2: History-based URL filtering.
Results

- Clustering
  - K=100 results

- Group Analysis
  - small num of groups have high peak values
  - most have small peak values
  - threshold = 0.45
  - 20 groups remain

Table 3: Clustering and group analysis results.

<table>
<thead>
<tr>
<th>Month</th>
<th>Total</th>
<th>Above threshold</th>
<th>Malicious</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sept 10</td>
<td>290</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>Jan 11</td>
<td>272</td>
<td>16</td>
<td>11</td>
</tr>
</tbody>
</table>

Figure 11: The distribution of peak values: percentage of pages sharing the same number of URLs within a group.
A New Attack

- Found another SEO attack that uses a different methodology
  - setting up SEO pages
  - boosting their page ranks
  - polluting the search index

- Differences b/t previous attack:
  - Not link to each other, rely on incoming links for page rank
  - Use cloaking
  - Extra level of redirection for exploit server
Study of SEO Attack

- Queries
  - Less than 5% of the top 500 Alexa Web sites ever submitted queries during the month of September 2010, while 46% of the compromised servers did.
  - Queries from the IPs of compromised servers are more frequent than those of legitimate sites.
- Matching Google and Bing queries

![Table 4: Matching Google and Bing search results using derived regular expressions.](image-url)