deSEO: Combating Search-Result Poisoning

John P John
Fang Yu, Yinglian Xie,
Arvind Krishnamurthy, Martin Abadi
University of Washington & MSR, Silicon Valley
The malware pipeline

- find vulnerable web servers
- compromise web servers and host malicious content
- spread malicious links via email, IM, search results
- bad stuff
The malware pipeline

- Malware links spread through:
  - spam emails, spam IMs, social networks, search results, etc.
- We look at *search results*
Is this really a problem?

- ~40% of popular searches contain at least one malicious link in top results
- Scareware fraud made $150 m. in profit last year
Is this really a problem?

- ~40% of popular searches contain at least one malicious link in top results
- Scareware fraud made $150m in profits last year
Contributions

• How does the search poisoning attack work?
  - examined a live attack involving 5,000 compromised sites

• What can we learn about such attacks?
  - identified common features in search poisoning attacks

• How can we defend against them?
  - developed deSEO, which detected new live SEO attacks on 1,000+ domains
Anatomy of SEO attack

- search engine
- compromised Web server
- redirection server
- exploit server
Anatomy of SEO attack

search engine

search query

compromised Web server

exploit server

redirection server
Anatomy of SEO attack

search engine

search query

compromised Web server

exploit server

redirection server
Anatomy of SEO attack

- Search engine
- Search query
- Compromised Web server
- Exploit server
- Redirection server
Anatomy of SEO attack

- Search engine
- Search query
- Compromised Web server
- Exploit server
- Redirection server
Anatomy of SEO attack

- search engine
- search query
- compromised Web server
- exploit server
- redirection server
Analysis of an attack

- Examine a specific attack
  - August - October 2010
  - 5,000 compromised domains
  - Tens of thousands of compromised keywords
  - Millions of SEO pages generated
How are servers compromised?

- Sites running osCommerce
- Unpatched vulnerabilities
- Allows attackers to host any file on the Web server - including executables

What files are uploaded?
What files are uploaded?

- php shell to manage file operations
What files are uploaded?

- php shell to manage file operations
- HTML templates, images
What files are uploaded?

- php shell to manage file operations
- HTML templates, images
- php script to generate SEO web pages
The main php script

www.example.com/images/page.php?page=kobayashi+arrested
The main php script
The main php script

www.example.com/images/page.php?page=kobayashi+arrested

- Obfuscated script
- Simple encryption using nested *evals*
The main script (de-obfuscated)

```php
<?php

global $hta;
if(!file_exists("./hta.cfg")) $hta = false;
else $hta = true;
@mkdir("./.news");
@chmod("./.news", 0777);

function crawl_page($url) {

function is_search_bots() {

function sendPage($keyword) {

function page404() {

function getRandom($key) {

function getNew($key) {

function loadTemplate($template) {

function getContentType($key) {

if ($_GET["q"])
    print sendPage($_GET["q"]);

elseif ($_GET["page"])
    print sendPage($_GET["page"]);
```
The main script (de-obfuscated)

<?php

global $hta;
if(!file_exists("./hta.cfg")) $hta = false;
else $hta = true;
@mkdir("./news");
@chmod("./news", 0777);

function crawl_page($url) {
}

function is_search_bots() {

function sendPage($keyword) {

function page404() {

function getRandom($key) {

function getNew($key) {

function loadTemplate($template) {

function getContent($key) {

if ($_GET["q"])
    print sendPage($_GET["q"]);

elseif ($_GET["page"])
    print sendPage($_GET["page"]);

}
The main script (de-obfuscated)

<?php

global $hta;
if(!file_exists("./hta.cfg")) $hta = false;
else $hta = true;
@mkdir("./news");
@chmod("./news", 0777);

function crawl_page($url) {
}

function is_search_bots() {
}

function sendPage($keyword) {
}

function page404() {
}

function getRandom($key) {
}

function getNew($key) {
}

function loadTemplate($template) {
}

function getContent($key) {
  if ($_GET["q"]){
    print sendPage($_GET["q"]);
  }
  elseif ($_GET["page"]){
    print sendPage($_GET["page"]);
  }
}
The main script (de-obfuscated)

```php
<?php

global $hta;
if(!file_exists("./hta.cfg")) $hta = false;
else $hta = true;
@mkdir("./news");
@chmod("./news", 0777);

function crawl_page($url) {

function is_search_bots() {

function sendPage($keyword) {

function page404() {

function getRandom($key) {

function getNew($key) {

function loadTemplate($template) {

function getContent($key) {

if ($GET["q"]){
    print sendPage($GET["q"]);
}
elseif ($GET["page"]){
    print sendPage($GET["page"]);
}
```

Check if search crawler
Generate page for keyword
Fetch:
  snippets from google
  images from bing
Add links to other compromised sites
The main script (de-obfuscated)

```php
<?php

    global $hta;
    if(!file_exists("./hta.cfg")) $hta = false;
    else $hta = true;
    @mkdir("./news");
    chmod("./news", 0777);

    function crawl_page($url) {
    
    function is_search_bots() {
    
    function sendPage($keyword) {
    
    function page404() {
    
    function getRandom($key) {
    
    function getNew($key) {
    
    function loadTemplate($template) {
    
    function getContent($key) {
    
    if ($GET["q"])
        print sendPage($GET["q"]);
    
    elseif ($GET["page"])
        print sendPage($GET["page"]);
    
    Check if search crawler
    Generate page for keyword
    Fetch:
    snippets from google
    images from bing
    Add links to other
    compromised sites
    Cache page
```
Dense link structure

- Other compromised domains found by crawling included links
- Each site linked to 200 other sites
- ~5,000 compromised domains identified
- Each site hosted 8,000 SEO pages
  - 40 million pages total
Poisoned keywords

- 20,000+ popular search terms poisoned
Poisoned keywords

- 20,000+ popular search terms poisoned
Poisoned keywords

- 20,000+ popular search terms poisoned
Poisoned keywords

- 20,000+ popular search terms poisoned
- Google Trends + Bing related searches
  - haiti earthquake
  - senate elections
  - veterans day 2010
  - halloween 2010
  - thanksgiving 2010...
Poisoned keywords

• 20,000+ popular search terms poisoned
• Google Trends + Bing related searches
  • haiti earthquake
  • senate elections
  • veterans day 2010
  • halloween 2010
  • thanksgiving 2010 ...
• 95% of Google Trends keywords poisoned
Redirection servers

- Three domains used for redirection
- Over 1,000 exploit URLs fetched
Redirection servers

- Three domains used for redirection
- Over 1,000 exploit URLs fetched

Almost 100,000 victims over 10 weeks
Evasive techniques

- Why can’t redirection behavior be easily detected?
  - Cloaking
  - Requiring user interaction
  - Redirection through javascript or flash
What are prominent features in search poisoning?

- Dense link structure
- Automatic generation of relevant pages
- Large number of pages with popular keywords
- Behavior of compromised sites
  - before - diverse content and behavior
  - after - similar content and behavior
What are prominent features in search poisoning?

- Dense link structure
- Automatic generation of relevant pages
- Large number of pages with popular keywords
- Behavior of compromised sites
  - before - diverse content and behavior
  - after - similar content and behavior
deSEO steps

1. History-based filtering
   select domains where many new pages are set up, different from older pages

2. Clustering suspicious domains
   using K-means++

3. Group similarity analysis
   select groups where new pages are similar across domains
Sample web URLs with trendy keywords

Sample web URLs with trendy keywords

History based detection
Sample web URLs with trendy keywords

History based detection

Domain clustering

- Lexical features of URLs

**String features** - keyword separators, arguments, filename, path

**Numerical features** - number of arguments, length of arguments, length of keywords

**Bag of words** - set of keywords
Sample web URLs with trendy keywords

History based detection

Domain clustering
- lexical features of URLs

Group analysis
- web page feature similarity
History based detection

Domain clustering
- lexical features of URLs

Group analysis
- web page feature similarity

Sample web URLs with trendy keywords
Sample web URLs with trendy keywords

- lexical features of URLs
Sample web URLs with trendy keywords

History based detection

Domain clustering
- lexical features of URLs

Group analysis
- web page feature similarity

Regular expressions
- to match URLs not in our sample

Sample web URLs with trendy keywords

History based detection

Domain clustering
- lexical features of URLs

Group analysis
- web page feature similarity

Regular expressions
- to match URLs not in our sample
deSEO findings

- 11 malicious groups from sampled web graph in January 2011
  - 957 domains
  - 15,482 URLs
- Revealed a new search poisoning attack
  - compromised Wordpress installations
  - cloaking to avoid detection
  - different link topology
Applying to search results

- 120 keyword searches in Google and Bing
- 163 malicious URLs detected in results
- 43 search terms affected
Conclusion

- Malware and SEO are big problems
- Analyzed an ongoing scareware campaign
  - Identified thousands of compromised domains
- Identified prominent features in SEO attacks and used them to build deSEO
  - Promising results on a partial dataset from Bing
  - Identified multiple live SEO attacks
Thank You

jjohn@cs.washington.edu