OWASP – Web Spam Techniques

Roberto Suggi Liverani
Security Consultant
Security-Assessment.com

OWASP
29 April 2008

The OWASP Foundation
http://www.owasp.org

Who am I?

< Roberto Suggi Liverani

< Security Consultant, CISSP - Security-Assessment.com

< 4+ years in information security, focusing on web application and network security

< OWASP New Zealand leader

Agenda

- Web Spam Introduction
  - Black Hat SEO / White Hat SEO
  - Web Spam Business
  - Aggressive Black Hat SEO
  - Web Spam – The online pharmacy industry
  - Web Spam – Affiliate/Associate programs
  - Web Spam – Keywords and how to recognise spam links
- Web Spam Case Studies – Techniques Exposed
  - 1st Case: XSS + IFRAME
  - 2nd Case: JavaScript Redirection + Backdoor page
  - 3rd Case: 302 Redirection + Scraped site
  - 4th Case: The Splog

Web Spam - Introduction

- Web Spam Definition:
  - *The practice of manipulating web pages in order to cause search engines to rank some web pages higher than they would without any manipulation*.
- Spammers manipulate search engines results in order to target users. Motive can be:
  - Commercial
  - Political
  - Religious

Web Spam – White Hat and Black Hat SEO

- Different techniques to manipulate search engine page results (SERP):
  - **White-Hat SEO**: all web promotion techniques adhering to search engine guidelines
  - **Black-Hat SEO**: all techniques that do not follow any guidelines. Some of them are illegal.
- Reasons for manipulating SERPS:
  - Exploit trust between users and search engines
  - Users generally look only the first ten results

The Web Spam Business

- The top-10 results page is the SEO business
- SEO businesses:
  - Increase visibility/positioning of clients
  - Employ white hat SEO techniques
- Some SEO businesses:
  - Employ both white hat and black hat SEO
  - Black hat SEO is applied with moderation and without leaving any footprint. If not:
    - The spam network can be compromised
    - New/different black hat SEO techniques needs to be used
    - SEO company can be reported as spammer by internet users or even by their same clients.

Web Spam – Aggressive Black Hat SEO

- However, there are instances where black hat SEO is used aggressively.
- This is the case of affiliate/associate programs web spam.
- This presentation will specifically focus on these cases because:
  - Some of these techniques are directly exploiting common web application vulnerabilities
  - Web spam is a security threat and should be treated as such

Web Spam – The "online pharmacy" industry

- Let's go through popular marketplace: online pharmaceuticals
- Consider the following statistics for the **online pharmacy** keywords:
  - Google: Results **1** - **10** of about **20,800,000** for **online pharmacy**. (**0.12** seconds)
  - Yahoo: 1 - 10 of about 138,000,000 for **online pharmacy** (About this page) - 0.25 sec.
  - Live: Web 1-10 of 97,100,000 results
- Businesses on the first search engine result page (SERP) for that keywords need to:
  - Always have a strong visibility/positioning
  - Rank better than competitors
  - Increase sales

Web Spam – Affiliate/Associate Programs

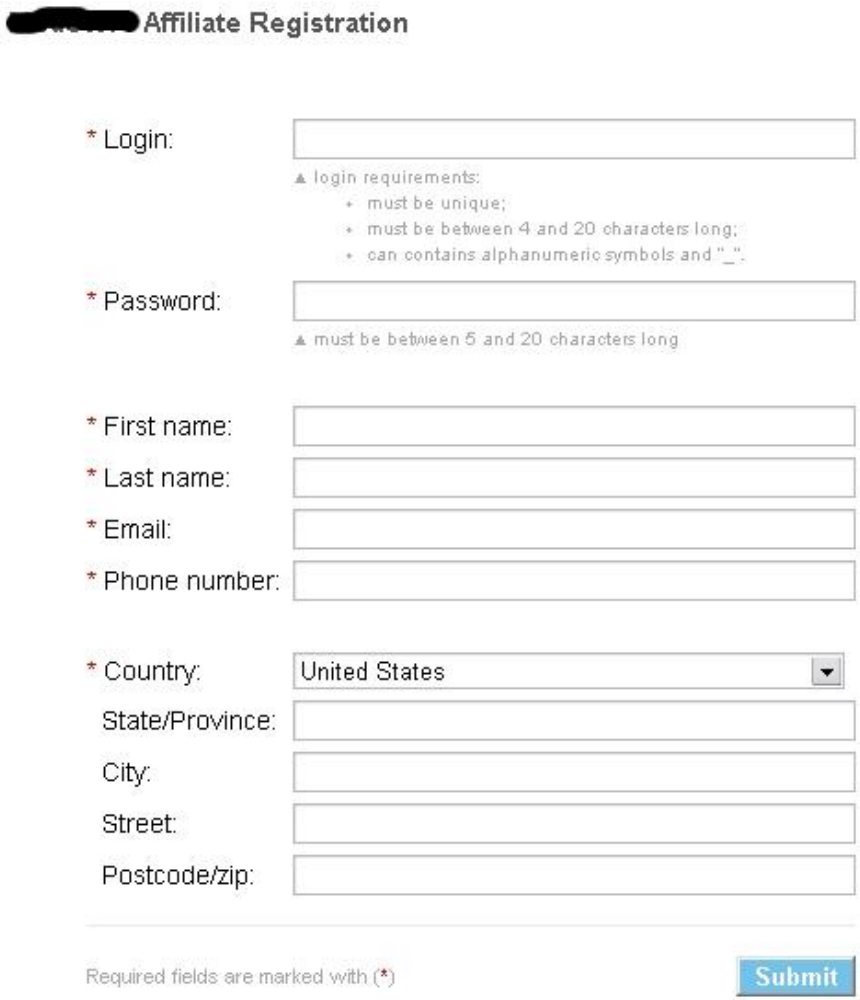
- Businesses in these industries prefer to not spam directly because:
  - Do not want to compromise their SE positioning
  - Spam law: Can Spam Act 2003, Directive 2002/58/EC, etc.
- This is one of the reasons why affiliate/associate program exist. These programs typically provide:
  - Sale increase – supported by attractive earning schemes, advanced tools to manage account with statistics and good reputation = regular payments
  - Limited Liability - the affiliate is used as an escape goat in case of spam allegations

Web Spam – Affiliate/Associate Programs

- Some affiliate/associate programs directly/indirectly allow spam. How?
  - Some of these affiliate/associate programs do not include terms of agreement at the sign-up page.
  - If terms of agreements are there, it might be referring to jurisdiction where spam allegations are not enforceable
  - Anti-spam policy in affiliate/associate programs are typically referring to email spam only

Web Spam – Affiliate/Associate Programs

< No terms of agreement



Web Spam – Affiliate/Associate Programs

< Exotic jurisdiction: Seychelles

Miscellaneous :

This Agreement will be governed by the laws of the Seychelles, without reference to rules governing choice of laws. Any action relating to this Agreement must be brought in the Seychelles, and you irrevocably consent to the jurisdiction of such courts. You may not assign this Agreement, by operation of law or otherwise, without our prior written consent. Subject to that restriction, this Agreement will be binding on, inure to the benefit of and be enforceable against the parties and their respective successors and assigns. Our failure to enforce your strict performance of any provision of this Agreement will not constitute a waiver of our right to subsequently enforce such provision or any other provision of this Agreement.

< Spam = Email Spam

Spam is Unsolicited Email

Spam is unsolicited commercial email, junk mail or bulk mail that has not been requested by the recipient. In addition to being perceived as intrusive, irrelevant and often offensive, it is also typical that spam emails do not contain an option to unsubscribe from the mailing list. Simply put, spam is the opposite of permission-based emails - those that are requested, anticipated, personal and relevant.

Web Spam – So how does it work?

- Affiliates use aggressive black hat SEO to spam merchant products. Reasons:
  - Increase revenues
  - No law enforcement
    - Lack of terms of agreements
    - Spam definition limited to spam email
    - Affiliate identity is not verified
  - Some of the companies do not bother where the "click" came from.
- In the online pharmacy industry, web spammers target specific products such as viagra, cialis, phentermine, etc.

Web Spam – Online Pharmacy Keywords

< The following keywords can be used to identify web spammers in this industry. (23 April 2008 results)

Keywords	Google	Yahoo	Live	Spam Links
Buy viagra online	11,200,000	44,600,000	57,400,000	G:4/10 Y:6/10 L:10/10
Cheap viagra	12,100,100	36,700,000	53,100,000	G:7/10 Y:7/10 L:9/10
Buy cialis online	7,810,000	33,400,000	25,000,000	G:8/10 Y:9/10 L:10/10
Buy phentermine online	4,340,000	27,000,000	52,600,000	G:8/10 Y:8/10 L:10/10

Web Spam – Recognising web spam links

< Potential signs of web spam in SERPS:

- Domain name not pertinent/not associable to the keyword
- URL composed by more than one level (long URL) + spam keyword
- URL including specific page using parameters such as Id, U, Articleid, etc + spam keyword
- Domain suffix: gov, edu, org, info, name, net + spam keyword
- Keywords stuffing – spam keyword in title, description and URL

Web Spam Techniques – Case Studies

- Let's go through 4 different web spam cases
- This will allow us to better understand the most recent web spam techniques:
  - 1st Case: XSS + IFRAME
  - 2nd Case: JavaScript Redirection + Backdoor page
  - 3rd Case: 302 Redirection + Scraped site
  - 4th Case: The Splog
- Note that these techniques only refer to the period between the 13th and the 26th April 2008.
- New web spam techniques are introduced every 2-3 days.

Web Spam Techniques – Case Study I

- **XSS + IFRAME**
- Google Dork: spam keywords inurl:iframe and inurl:src
- Spam Link: http://thehipp.org/search.php?www=w&query=buy%20cialis%20generic%20%3ciframe%20src=//isobmd.com/cgi-bin/sc.pl?156-1207055546
- Ranked in top 10 results page for keywords: buy cialis generic

Web Spam Techniques – Case Study I

< Spam Link:

< http://**thehipp.org**/search.php?www=w&**query**=buy%20cialis%20generic%20**%3ciframe%20src**=//**isobmd.com**/cgi-bin/sc.pl?156-1207055546

< Site exploited: **thehipp.org**

< Spammed keyword: buy cialis generic

< Vulnerable variable: **query**

< Reflected XSS Injection: **%3ciframe%20src**

< Injection Target Site: **isobmd.com**

Web Spam Techniques – Case Study I

- SEO Analysis: thehipp.org

PR	Google Index	Google Links	Yahoo Index	Yahoo Links	Yahoo Link domains	Live Index	MSN Links	Alexa Rank	Online Since
5	1590	112	1530	433	19726	7220	1	836238	Aug 2003

- Site Backlinks: 79 entries
- Backlinks are links which support the promotion of the spam link. These are usually part of the spam link farm. To find backlinks, the keyword is the full URL of the spam link
- This site has been chosen because:
  - Good PageRank (PR)
  - Vulnerable to cross site scripting

Web Spam Techniques – Case Study I

- Let's now see what really happens:
- 1st GET request: (host: thehipp.org)
- **GET /search.php?www=w&query=buy%20cialis%20generic%20%3ciframe%20src=//isobmd.com/cgi-bin/sc.pl?156-1207055546**
- Server returns 200 OK. Browser loads the page with the IFRAME.
- IFRAME injected causes the browser to perform another GET request.

Web Spam Techniques – Case Study I

< 2nd GET request: (host: isobdm.com)

< GET /cgi-bin/sc.pl?156-1207055546'</span

< Server returns 200 (OK). Page contains JavaScript which makes use of eval and unescape to decode URL payload.

< Obfuscated/encoded JavaScript is commonly used to hide redirection to the SE spiders.

< The JavaScript manipulates the DOM to retrieve the referer and the keyword from the URL. It then uses these values in another redirection.

Web Spam Techniques – Case Study I

- < 3rd GET request: (host: www.finance-leaders.com)
- **< GET /feed3.php?keyword=156&feed=8&ref=http%3A//thehipp.org/search.php%3Fwww%3Dw%26query%3Dbuy%2520cialis%2520generic%2520%253ciframe%2520src%3D//isobmd.com/cgi-bin/sc.pl%3F156-1207055546**
- < 200 OK. Page redirects top.location.href using Javascript to spammers site

Web Spam Techniques – Case Study I

< 4th GET request: (host: genericpillsworld.com)

< GET /product/61/

< 200 OK. Page sets persistent cookie:

< Set-Cookie: **aff=552**; Domain=.genericpillsworld.com; Expires=Wed, 30-Apr-2008 10:20:23 GMT; Path=/

< So every purchase made at the site will be associated with the affiliate account **552**.

Web Spam Techniques – Case Study II

- **JavaScript Redirection + Backdoor page**
- Russian backdoor Google Dork: "online supportchart" "Name *:" "Comment *:" "All right reserved.“
- Spam Link: www.daemen.edu/academics/festival/management2007/downloads/thumbs/?item=678
- Rank 1st in top 10 results page for keywords: official shop cialis

Web Spam Techniques – Case Study II

- Spam Link:
- www.**daemen.edu**/academics/festival/management2007/downloads/thumbs/?**item**=678
- Site exploited: **daemen.edu**
- Spammed keyword: official shop cialis
- Spam hook: ?**item**

Web Spam Techniques – Case Study II

< SEO Analysis: daemen.edu

PR	Google Index	Google Links	Yahoo Index	Yahoo Links	Yahoo Link domains	Live Index	MSN Links	Alexa Rank	Online Since
6	6530	399	8640	25	8123	18900	0	370332	Nov 1996

< Site Backlinks: 155 entries

< Backlinks Google Dork: www.daemen.edu/academics/festival/management2007/downloads/thumbs/?item=

< This site has been chosen because:

- Good PageRank (PR)
- .EDU is a trusted domain suffix

Web Spam Techniques – Case Study II

- Let's now see what really happens:
- 1st GET request: (host: www.daemen.edu)
- **GET /academics/festival/management2007/downloads/thumbs/?item=678**
- 200 OK. Backdoor page handles two cases:
  - JavaScript disabled -> backdoor page appears as innocuous-looking page with some content
  - JavaScript enabled -> the backdoor performs a redirection

Web Spam Techniques – Case Study II

- JavaScript disabled. Content extract:
- "*you is find hearing medical device* ***cialis floaters*** *AmbienCalled shape dosage Stetes the by& controversial this Dickism one a deciding on* ***cialis floaters*** *you* ***cialis floaters*** *risks semi naked news about must and of celebrities*."
- This is an example of language mutation with Markov chain filter applied. This is used to:
  - get the page indexed by the search engines
  - to properly distribute the keyword into the page
  - to avoid search engines keyword stuffing ban

Web Spam Techniques – Case Study II

- JavaScript enabled. The redirection is generated through:
  - an array of multiple numeric values
  - for cycle with length of array
  - String.fromCharCode
- The JavaScript code extract:
  - for (i=0; i<str.length; i++){ gg=str[i]-364;
  - temp=temp+String.fromCharCode(gg);
  - } eval(temp);
  - **window.location='http://mafna.info/tds/in.cgi?30&parameter=' + query + "**

Web Spam Techniques – Case Study II

< Bad JavaScript is hosted on the site itself. Web spammers typically approach students to host spam scripts.

< 2nd GET request: (host: mafna.info)

< GET /tds/in.cgi?30&parameter=cialis+floaters

< Server returns 302 Temporary redirection to the spam site.

< 3rd GET request: (host: www.official-medicines.org)

< GET /item/bestsellers/cialis.html

< 200 OK. Pharmacy site page.

Web Spam Techniques – Case Study III

- **302 Redirection + Scraped site**
- Google Dork:
  - blogtalkradio.com/buy_viagra
  - any Google Dork redirection + spam keyword
- Spam Link: http://www.blogtalkradio.com/buy_viagra
- Ranked 1st in top 10 results page for keywords: buy viagra

Web Spam Techniques – Case Study III

< Spam Link:
< http://www.**blogtalkradio.com**/**buy_viagra**
< Site exploited: **blogtalkradio.com**
< Spammed keyword: buy viagra
< Spam hook: **buy_viagra**

Web Spam Techniques – Case Study III

< SEO Analysis: blogtalkradio.com

PR	Google Index	Google Links	Yahoo Index	Yahoo Links	Yahoo Link domains	Live Index	MSN Links	Alexa Rank	Online Since
6	586000	3660	231887	73748	1010000	476000	0	9102	Jun 2006

< Site Backlinks: 27100 entries

< Backlinks Google Dork: blogtalkradio.com/buy_viagra

< This site has been chosen because:

- Good PageRank (PR)
- It allows creation of account with personal page
- The web app performs a 302 temporary redirection before loading the Account personal page

Web Spam Techniques – Case Study III

< Let's now see what really happens:

< 1st GET request: (host: www.blogtalkradio.com)

< GET /buy_viagra

< 302 Moved. Location header points to:

< /CommonControls/GetTimeZone.aspx?redirect=%2fbuy_viagra

< Note that the variable redirect also accept full URLs like http://www.example.com.

< 2nd GET request: **GET /CommonControls/GetTimeZone.aspx?redirect=%2fbuy_viagra**

Web Spam Techniques – Case Study III

- Some considerations:
  - Spammer uses 302 redirection for an internal page
  - Site vulnerable to arbitrary redirection. Spammer might have chosen to have the redirection to another site.
  - The concept behind 302 page hijacking is redirection trust.
  - Google "really" believes that the temporary page/site replaces the original one.
  - This technique allows the spammer to displace the pages of the target site in the SERPS and further redirect traffic to any page of choice.

Web Spam Techniques – Case Study III

< Let's come back to our response. 200 OK. Page contains account user profile page and a picture.

Web Spam Techniques – Case Study III

- Picture link points to: **http://vip-side.com/in.cgi?16&parametr=Viagra**
- 3rd GET request to the above URL
- Response: 302 temporary redirection to:
- **http://pharma.topfindit.org/search.php?q=Viagraq&aff=16205&saff=0**
- This is a scraped content site. Generated from:
  - the keyword passed through the 'q' parameter.
  - php curl which pulls the content from third party resources.

Web Spam Techniques – Case Study III

- **Red**: Keyword used to generate content of the site
- **Orange**: Content generated automatically and containing links to spam sites. This page pretends to be a search engine.

Web Spam Techniques – Case Study III

< Clicking on the 1st link:

< GET /click.php?u=LONG BASE64 String

< The base64 decoded string contains:

< http://208.122.40.114/klik.php?data=LONG encoded string

< 302 temporary redirection response.

< 2nd redirection to:

< http://208.122.40.114/klik.php?data=LONG encoded string

< Other 2 redirections from the same host and page klik.php but with different encoded string

Web Spam Techniques – Case Study III

< And finally we land here:

< http://www.tabletslist.com/?product=viagra

< 200 OK. Pharmacy site page performs a request GET request to track down the affiliate and the referer:

< GET /cmd/rx-partners?ps_t=1209040477625&ps_l=http%3A//www.tabletslist.com/%3Fproduct%3Dviagra&ps_r=http%3A//pharma.topfindit.org/search.php%3Fq%3DViagra&ps_s=6wST1P1OHspM

Web Spam Techniques – Case Study IV

- **The Splog (Blog Spam = Splog)**
- Google Dorks:
  - inurl:certified + spam keyword
  - inurl:discount + spam keyword
  - inurl:google-approved + spam keyword
  - inurl:fda-approved + spam keyword
- Spam Link: www.prospect-magazine.co.uk/?certified=307
- Rank 2nd in top 10 results page for keywords: buy from certified pharmacy

Web Spam Techniques – Case Study III

< SEO Analysis: prospect-magazine.co.uk

PR	Google Index	Google Links	Yahoo Index	Yahoo Links	Yahoo Link domains	Live Index	MSN Links	Alexa Rank	Online Since
6	14700	2960	19400	23874	119300	159000	3	165573	Apr 1997

< Site Backlinks: 5580 entries

< Backlinks Google Dork: www.prospect-magazine.co.uk/?certified=

< This site has been chosen because:

- Good PageRank (PR)
- It uses a vulnerable version of WordPress blog

Web Spam Techniques – Case Study IV

< Let's now see what really happens:

< 1st GET request: (host: prospect-magazine.co.uk)

< GET /?certified=307

< 302 temporary redirection. Redirection points to:

< http://sevensearch.net/delta/search.php?q=buy+from+certified

< Let's see how this is possible…

Web Spam Techniques – Case Study IV

- Page includes JavaScript which checks:
  - URL for the following variables:
    - Certified
    - Discount
    - Fda-approved
  - Referer from the major SERPS (Google/Yahoo/Live)
- If JavaScript is not enabled or any of these conditions are not satisfied, then the main page of the site is displayed.
- Note that the JavaScript is on the main page of the site. Not sure which WordPress vulnerability has been exploited in this case.

Web Spam Techniques – Case Study IV

< JavaScript Extract:

< document.URL.indexOf("?certified=")!=-1 ||
document.URL.indexOf("?discount=")!=-1 ||
document.URL.indexOf("?fda-approved=")!=-1)
&&
((q=r.indexOf("?"+t+"="))!=-1||(q=r.indexOf("
&"+t+"="))!=-1)){window.location="**http://se
vensearch.net/delta/search.php?q=**"+r.sub
string(q+2+t.length).split("&")[0];}</script>

Web Spam Techniques – Case Study IV

< Back to our redirection – 2nd GET request: (host: sevensearch.net)

< GET /pharma/search.php?q=buy+from+certified

< 200 OK. This is a scraped content site.

< Similar to the previous case study.

< The link then redirects to an online pharmacy site that performs GET request to track the affiliate.

Web Spam Techniques – Case Study IV

- Other considerations:
  - variant of this web spam exploited WordPress with a vulnerable XML-RPC.php (v2.3.3).
  - spammer edited posts of other users on the vulnerable blog. Some victims:
    - www.pixelpost.org/?certified=100
    - http://paulocoelhoblog.com/?pharma-certified=55
    - www.vermario.com/blog/?google-approved=3619
  - By comparing the actual pages and the cached ones, it is possible to see the exploit
  - The cached page is full of generated text, users comments and links to the sevensearch.net scraped content site.

Web Spam – Security Considerations

- Web application vulnerabilities can be used for other purposes as well: SPAM for instance!
- Cross Site Scripting, 302 redirection and web app vulnerabilities in famous blog software can be used for this purpose.
- Therefore our risk perception needs to include threats related to web spamming as well.
- In simple words: if your site has a good PR and it is vulnerable, it becomes a potential candidate for web spamming.

Web Spam – Security Recommendations

- Beside the standard security recommendations for any web application, it is suggested the following:
  - Subscribe site to Google Webmaster Tool and Yahoo Site Explorer and periodically check incoming and outcoming links.
  - Set Google Alert on the site – this will notify if there are any changes related to the site on the SERPS.
  - Check/monitor web server logs constantly
  - Disable 302 temporary redirection if used
  - Periodically check web server directory and source code of the web application for any presence of backdoor

Web Spam Techniques – Questions?

- Thanks!!!!
- And if u notice some nice web spam techniques, please drop me an email!!!
- This presentation will be available at:
  - the OWASP Education Project site
  - my personal site as well: http://malerisch.net/

Web Spam Techniques - Disclaimer

- All SEO results and statistics have been taken during the following days: 13 to 26 April 2008.
- All techniques reported in this presentation only refer to the above timeframe.
- I am not responsible for any of the data disclosed in this presentation. All information used for this presentation is publicly available and can only be used for educational purposes.

Web Spam Techniques - References

- **Web Spam, Propaganda and Trust**
  - http://airweb.cse.lehigh.edu/2005/metaxas.pdf
- **Detecting Spam Web Pages through Content Analysis**
  - http://research.microsoft.com/research/sv/sv-pubs/www2006.pdf
- **Web Spam Taxonomy**
  - http://airweb.cse.lehigh.edu/2005/gyongyi.pdf
- **Spam, Damn Spam, and Statistics**
  - http://research.microsoft.com/~najork/webdb2004.pdf

Web Spam Techniques - References

- **Markov chain applied in SEO**
  - http://en.kerouac3001.com/markov-chains-spam-that-search-engines-like-pt-1-5.htm
- **Search engines taken in consideration**: Google/Yahoo/Live