Adversarial Information Retrieval on the Web

Marcin Kaminski

Motivation

- Adversarial Information Retrieval has been playing a big role in nowadays Web oriented environment. Almost everybody stores some sensitive content online. Therefore, it is important to know how Adversarial IR can affect us so we can be prepared and successfully protect our confidential data.

Introduction

During 14th International World Wide Web Conference in Chiba [1], Japan and SIGIR 2006 Conference in Seattle[2] the following subjects on Adversarial Information Retrieval on the Web were discussed:

- search engine spam and optimization,
- crawling the web without detection,
- link-bombing (a.k.a. Google-bombing),
- comment spam, referrer spam,
- blog spam (splogs),
- malicious tagging,
- reverse engineering of ranking algorithms,
- advertisement blocking, and
- web content filtering

Introduction (cont)

- Using web crawlers to steal confidential information.

Web Crawlers

- “A web crawler (also known as a Web spider or Web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner. Other less frequently used names for Web crawlers are ants, automatic indexers, bots, and worms” [3].

Web Crawlers (cont)

- a software agent
- populates the IP addresses on the fly or starts with the predefined list of URL addresses to visit. As it visits these URLs, it gathers all the links it has identified and adds them to the list of URLs to visit, called the crawl frontier . It will then visit these newly added URLs recursively. **However, this flow allows some malicious action.**
- a web crawler may reach a sensitive web content which originally was not intended to be available for public view.
- the art of querying search engines in order to retrieve sensitive information gathered by web crawlers.

Google Hacking

- the art of creating complex search queries in order to obtain sensitive information such as credit card numbers, social security numbers, and passwords.
- Google operators are used to locate desired strings of text within search results. These strings could be i.e. “user name”, “password”, “SSN” etc. Similar sensitive information can also be retrieved using Google Code Search to search for insecure coding practices [4].
- Also other search engines like Yahoo! or MSN can be used to search for sensitive information.

Retrieval of Sensitive Information

- The idea of web crawling and “Google Hacking” was combined together by two employees of Estonian financial services firm Lohmus Haavel & Viisemann, Oliver Peek and Kristjan Lepik. Their case found its end in the court.

Lohmus Haavel & Viisemann case

- In November 2005, the U.S. Securities and Exchange Commission (SEC) has accused an Estonian financial firm Lohmus Haavel & Viisemann and two of its employees Oliver Peek and Kristjan Lepik , of carrying out a fraudulent hacking scheme that netted them at least $7.8 million.
- The agency accused them of using a web crawler to steal information related to more than 360 embargoed press releases in advance of their official distribution time from news and press release website Business Wire (BW).
- A statement from the SEC claims the stolen information allowed the Lohmuss to place their trades ahead of the release time of news involving mergers, earnings and regulatory action [7].

How the scheme could actually work so well?

Information on BW's servers was secured and blocked from being indexed by any search engine prior to the release time. Therefore, different technique than "Google hacking" had to be used to obtain the sensitive information.

- In June 2004 Lohmus became a client of BW. It is now allowed to publish his press release through the BW's system and has access to the secure client website of BW.
- April 2005 – October 2005, Using a web crawler, they did gain unauthorized access to the confidential information contained pending press releases of other BW clients including the expected time of issuance. The retrieval of such confidential information was possible due to the flow in the BW security. That is when a client submits a material for the news release it is lined up for public dissemination, either immediately or at its designated time. Now the Lohmus' crawler gets into action. Its access it's Business Wire private content and systematically goes through the underlying content of other clients, retrieving all the information it needs.
- Advanced techniques used to guess the links under other's client data may reside.
- Once the information is found it is send back to the Lohmuss' computer.

How it ended…

- Retrieved data allowed Lohmus to execute hundreds buy/sell trades. Sometimes trades were executed within 5 minutes the data was retrieved [5].
- Lohmus netted over $7.8 millions on those trades within a period of seven months.
- Therefore, what could be initially thought as just another example of web crawler based Adversarial IR actually turned into significant material gains/loses for the parties involved.

Conclusion

- What is disturbing here is that Lohmus was almost a 10 years old investment bank with offices in several countries. It did not, however, stopped him from performing an Adversarial Information Retrieval and breaking a law. It should be a warning to all financial institutions to pay a biggest attention to the security issues.
- Retrieval can turn to be very dangerous if aimed at the weak security. Keeping the sensitive data offline would be a good way to avoid troubles.

Future Work

- Definitely more studies on sophisticated Adversarial IR techniques should be done to protect sensitive data.
- Limit accessibility of sensitive data from the Internet. For sure, sure that it is not the case anymore at Business Wire.

References

- [1] 14th International World Wide Web Conference in Chiba , http://www2005.org/
- [2] SIGIR 2006 Conference in Seattle, WA http://www.sigir2006.org/
- [3] Kobayashi, M. and Takeda, K. (2000). *Information retrieval on the web. ACM*
- *Computing Surveys* 32 (2): 144-173.
- [4] Johnny Long (Author), Ed Skoudis, Alrik van Eijkelenborg, *Google Hacking for*
- *Penetration Testers*
- [5]Lohmus Haavel & Viisemann, et al.: Lit. Rel. No. 19450 / November 1, 2005,
- http://www.sec.gov/litigation/litreleases/lr19450.htm
- [6] Cyber Crimes and Solutions. http://www.selfseo.com/story-17135.php
- [7] SEC accuses Estonian firm of financial news hack | CNET News.com.
- http://news.com.com/SEC+accuses+Estonian+firm+of+financial+news+hack/2100- 7348_3-5931168.html