Google Hacking

Information Security Summit
Cleveland, Ohio

Pete Garvin
pgarvin@protectus.com
October 2005
Google Hacking Overview

• A few words about Google
• What is Google Hacking?
• Why it’s relevant
• How-to
• Defenses
• References
Google Overview

Googlebot

Web Servers (robots.txt) (meta-tags)

Google Database

Google Hacking - October 2005
Did Google get hacked?

No it didn’t, it just has lots of options. Select Preferences from the main page.
What is Google Hacking?

• The technique of using search engines to:
  – Find vulnerable targets
    • Misconfigured servers
    • Web based admin interfaces
    • Servers running a particular version of software
  – Find sensitive data
    • “Unpublished” web pages
    • Directory listings
    • Databases
What is Google Hacking? (cont)

• The technique of using search engines to:
  – Harvest data
    • Email addresses
    • Names and postal addresses
    • Server names
    • UserIDs and passwords
  – Perform reconnaissance
    • Find servers without scanning
    • Find relationships between web sites
Why its relevant

- Very powerful
- Growing in popularity
- Can be done anonymously
- Web based interfaces are everywhere
- Web servers are everywhere
- Not limited to Google
- Also includes listserv’s / forums
How-to

• Manually
  – Type a search phrase
  – Use the Google Hacking Database

• Page parsing
  – Use script to collect & parse desired data
  – Violates Google’s Terms of Service (TOS)

• Google API
  – Use a program to interact with the search engine (WSDL & SOAP)
Search Operators

- intitle – finds pages with phrase in title
- inurl – finds pages with phrase in the URL
- filetype – finds specific types of files
- cache – displays cached version of page
- site – narrows search to specific sites
- link – searches for links to a page
- Many others ……..see references
Example 1:
"Copyright (c) Tektronix, Inc." "printer status"
Example 2:
inurl:"ViewerFrame?Mode="
Example 2: Looking in the other direction...
Notes on Examples 1 & 2 ......

- Edit the URL directly – it’s faster

- Many things have a web admin interface
  - Copiers / Printers
  - Cameras
  - Firewalls / Network gear / Security appliances
  - Applications
Example 3:
intitle:index.of site:ibm.com
Example 3 (cont):
Select the Parent Directory

Directory Browsing Prohibited.

Good Job
IBM!
Example 4: userid birthdate filetype:csv
Notes on Examples 3 & 4…….

• We can find:
  – databases……..filetype:mdb
  – config files……..filetype:conf
  – mailboxes……..filetype:pst
  – CSV files……..filetype:csv
  – spreadsheets….filetype:xls

• Partial list of filetypes available at:
  http://www.google.com/help/faq_filetypes.html
Example 5: "Apache/1.3" intitle:index.of

Not all hits are vulnerable servers
Example 6: "cache:hurricane"
Notes on Examples 5 & 6 ……

• Unexpected results sometimes occur
• Not all hits are vulnerable servers
• Need to use search reduction to refine the results
• Some operators can be combined
• Other operators are used alone
How else could this be used?

• Create a computer worm to search Google for vulnerable targets
• Santy worm
  – December 2004
  – allinurl: viewtopic.php topic=12345
• MyDoom-O
  – July 2004
  – Harvested email addresses from Google
Example 7:
allinurl: viewtopic.php topic=12345
Example 8:

```
intitle:"PHP Shell "*" "Enable stderr" filetype:php
```
Google Hacking Database

• Point & click Google hacking!

• Categories include:
  – Advisories and Vulnerabilities
  – Error Messages
  – Files containing usernames & passwords
  – Pages containing login portals
  – Pages containing network or vulnerability data
  – Sensitive Directories & Online shipping info
  – Various Online Devices
  – Vulnerable Files & Servers
  – Web Server Detection
Remaining Anonymous

• Default cache behavior
  – Load external references from the original web page

• To remain anonymous
  – Use an anonymous http proxy
  – Append the strip=1 parameter
    • On the URL of a page cached by Google
    • Serves only the text of the website
Cached Pages

1. Cached pages (text)

2. Images

Google Database

Target web server
Automation

• Page Scraping or Page Parsing
  – Directly from the command line
  – Gooscan – automates queries
  – Both violate Google’s Terms Of Service (TOS)

• Google API
  – BiLE – Bi-directional Link Extractor
  – finds relationships between websites
Page Parsing
From a Command Line

- Google Database
  - Lynx
  - http://

Text file
- sed | awk | sort
  - http://server1.ibm.com
  - http://server2.ibm.com
  - .......... http://server98.ibm.com

- Remember:
  - This violates Google’s TOS
  - One server can have many names

Google Hacking - October 2005
Defenses

• Good practices and common sense go a long way
  – If it doesn’t need to be on your web servers, don’t put it there
    • Internal
    • External / public
  – Use web server config / security checklists
  – Only resolve hosts that require public naming
Defenses

• Google Hack your site
  – You might be surprised :-o
  – If you don’t, someone else will!

• Use a robots.txt file
  – Two edged sword
  – Respectable search engines use it but….­
  – Observance of robots.txt is optional
Defenses

• Disable directory browsing on web servers
  – If you must display contents of a directory, disable directory browsing

• Metatags
  – NOARCHIVE
  – NOSNIPPET
  – NOINDEX, NOFOLLOW

• Password protection
Defenses

• What if Google grabs some sensitive information from your site?
  – Remove the data from your site
  – Determine the source of the leak
  – Refer to www.google.com/remove.html
    • Update web site & wait for Googlebot
    • Use an automatic URL removal system
    • Takes hours to days depending on method used
Defenses

Introduction

GHH is the reaction to a new type of malicious web traffic: search engine hackers. GHH is a "Google Hack" honeypot. It is designed to provide reconnaissance against attackers that use search engines as a hacking tool against your resources. GHH implements honeypot theory to provide additional security to your web presence.

Google has developed a powerful tool. The search engine that Google has implemented allows for searching on an immense amount of information. The Google index has swelled past 8 billion pages (February 2005) and continues to grow daily. Mirroring the growth of the Google index, the spread of web-based applications such as message boards and remote administrative tools has resulted in an increase in the number of misconfigured and vulnerable web apps available on the Internet.

These insecure tools, when combined with the power of a search engine and index which Google provides, results in a convenient attack vector for malicious users. GHH is a tool to combat this threat.

GHH emulates a vulnerable web application by allowing itself to be indexed by search engines. It's hidden from casual page viewers, but is found through the use of a crawler or search engine. It does this through the use of a transparent link which isn't detected by casual browsing but is found when a search engine crawler indexes a site. The transparent link (when well-crafted) will reduce false positives and avoid a fingerprint of the honeypot.

The honeypot connects to a configuration file, and the configuration file writes to a log file which is chosen during configuration. The log file contains...
References

- www.google.com/help/faq_filetypes.html
- filext.org
- johnny.ihackstuff.com
- www.google.com/apis
- www.google.com/terms_of_service.html
- www.sans.org/GoogleCheatSheet.pdf
- www.robotstxt.org
- www.google.com/remove.html
- www/google.com/webmasters
Any questions?