Google Hacking Against Privacy

Emin Islam Tatlı

University of Mannheim, Department of Computer Science
tatli@th.informatik.uni-mannheim.de
(on leave to the University of Weimar)

Key words: Google hacking, privacy, privacy searches

Summary. Google facilitates our lives by finding any searched information within a single-click time. On the other hand, Google threatens our privacy by revealing our personal data to others. In this paper, we give examples of Google hacking against user privacy and discuss the countermeasures to protect our privacy from Google or in general from the search engines.

1 Motivation

Google is the most popular web search engine in internet. It indexes any information from web servers thanks to its hardworking web crawlers. But many personal data that should be kept secret and confidential are indexed by Google, too. This threatens our privacy. Personal data like name, address, phone numbers, emails, CVs, chat logs, forum and mailing list postings, username-password pairs for login sites, private directories, documents, images, online devices like web cameras without any access control, secret keys, private keys, encrypted messages, etc. are all available to others via Google. In addition to privacy risks, there are many more other security risks that can be revealed by Google. There exist an online database [3] which contains 1423 different Google hacking search queries by June 2007.

This paper focuses on the advanced search queries that illustrate how Google reveals personal private data which are expected to stay private. Actually, other search engines reveal such private data too. But Google has become our target due to its huge index size. The paper is organized as follows: Section 2 summarizes the useful parameters for the advanced search in Google. In Section 3, examples of search queries against privacy are given. Section 4 explains possible security measures against Google privacy hacking. Finally, Section 5 explains our future work plan.

2 Google Search Parameters

In addition to the basic search operators (i.e. +,-,.), Google supports more parameters for the advanced searches and filters its results according to the parameters provided by the user.

The *[all]inurl* parameter is used to filter out the results according to if the given url contains a certain keyword or not. If more keywords are needed to filter, the *allinurl* parameter should be used. *[all]intitle* filters the results according to the title of web pages. *[all]intext* searches the keywords in the body of web pages. With the parameter *site* you can do host-specific search. *filetype* and *ext* parameters have the same functionality and are needed to filter out the results based on the file extensions like html, php, pdf, doc, etc. The minus sign (-) can be put before any advanced parameter and reverses its behavior. As an example, a search query containing the parameter *-site:www.example.com* will not list the results from www.example.com. The operator `"|"` or the keyword OR can be used for binding different searches with the *logical OR* operation.

3 Privacy Searches

Google can reveal many personal data when its advanced search parameters are used. We have grouped private data searches into four different sections according to the privacy level. These are *identification* data, *sensitive* data, *confidential* data and *secret* data searches.

3.1 Identification Data

The identification data is related to the personal identity of users. Name, surname, address, phone number, marital status, CV, aliases, nicknames used over internet, etc. are the typical examples of identification data. Some private data searches would focus on a certain person and we choose the name "Thomas Fischer" which is a very common personal name in Germany.

Name, Address, Phone, etc.

You can search web pages and documents which contain keywords like name, address, phone, email, etc., optionally for a certain person or within certain document types.

```
allintext:name email phone address intext:"thomas fischer" ext:pdf
```

Twiki [1] is a wiki-based web application that is commonly used for project management purpose. Inside TWiki, user data like name, address, phone numbers, web pages, location, emails, etc. are stored. If the required authentication techniques are not applied, unauthorized people can also access this data.

```
intitle:Twiki inurl:view/Main "thomas fischer"
```

In addition to Google search, other search engines with the capability of people finder can also be very helpful for gaining identification data. Yahoo's People Search [2], Lycos's WhoWhere People Search [3] or eMailman's People Search [4] connecting public ldap servers are examples of such services.

Curriculum Vitae

You can search for the keyword CV (curriculum vitae) that contain many personal data. This search can be extended by including translations of the CV in different languages. For example, Lebenslauf can be integrated in the search as the german translation for CV.

```
intitle:CV OR intitle:Lebenslauf "thomas fischer"
intitle:CV OR intitle:Lebenslauf ext:pdf OR ext:doc
```

Usernames

Webalizer web application [5] collects statistical information of web sites about visitor activities. The most commonly used login usernames are also stored by Webalizer.

```
intitle:"Usage Statistics for" intext:"Total Unique Usernames"
```

3.2 Sensitive Data

The sensitive data means the data which are normally public but may contain private personal data and whose reveal may disturb its owner. The examples are emails, postings sent to lists, sensitive directories and Web2.0 based applications.

[1] Twiki: http://twiki.org
[2] Yahoo People Search: http://people.yahoo.com
[3] Lycos People Search: http://peoplesearch.lycos.com
[4] eMailman People Search: http://www.emailman.com/ldap/public.html
[5] Webalizer: http://www.mrunix.net/webalizer/

Forum Postings, Mailinglists

PhpBB [6] is a widespread web forum application. It enables to find out all postings sent by a particular user. The following search finds out all postings sent with the alias thomas to different phpBB-based forums.

```
inurl:"search.php?search_author=thomas"
```

Mailman [7] is a well-known mailing list manager. The following search gives all email postings which are sent to mailman-based lists and related to *Thomas Fischer*.

```
inurl:pipermail "thomas fischer"
```

Sensitive Directories

Backup directories can contain also some sensitive data about users, organizations, companies, etc.

```
intitle:"index of" inurl:/backup
```

Web2.0

The next generation internet Web2.0 introduces more privacy risks. With Web2.0, people share more personal data with others. The following searches are based on the favorite Web2.0 sites like Yahoo's Image Sharing [8], Google's Blogger [9] and Google's Video Sharing [10]. Instead of searching through Google, searching directly on the original sites would give more efficient results.

```
"Thomas Fischer" site:blogspot.com
"thomas" site:flickr.com
"thomas" site:youtube.com
```

3.3 Confidential Data

The confidential data is normally expected to be non-public for others except for a group of certain persons, but with Google it becomes possible to access to such private data as well.

[6] PhpBB Forum: http://www.phpbb.com
[7] Mailman List Manager: http://www.gnu.org/software/mailman/
[8] Yahoo Image Sharing: http://www.flickr.com
[9] Google's Blogger: http://www.blogspot.com
[10] Google Video Sharing: http://www.youtube.com

Chat Logs

You can search for chat log files related to a certain nickname.

```
"session start" "session ident" thomas ext:txt
```

Username and Password

Username and password pairs can be searched within sql dump files and other documents.

```
"create table" "insert into" "pass|passwd|password" (ext:sql |
ext:dump | ext:dmp | ext:txt)
"your password is *" (ext:csv | ext:doc | ext:txt)
```

Private Emails

Microsoft Outlook and Outlook Express store the personal emails within single database files like incoming messages within inbox.dbx. The following searches target the email storage files stored by Outlook Express or Microsoft Outlook.

```
"index of" inbox.dbx
"To parent directory" inurl:"Identities"
```

Confidential Directories and Files

Confidential directories and files can be revealed with the following query.

```
"index of" (private | privat | secure | geheim | gizli)
```

In order to prevent web crawlers to list private directories, Robot Exclusion Standard [6] is used. But it also enumerates a number of private directory paths within world-readable robots.txt files.

```
inurl:"robots.txt" "User-agent" ext:txt
```

Not only directories but also private documents and images can be searched through Google.

```
"This document is private | confidential | secret" ext:doc |
ext:pdf | ext:xls
intitle:"index of" "jpg | png | bmp" inurl:personal | inurl:private
```

Online Webcams

Online webcameras come along with their software for the remote management over internet. Based on the type of the webcam, you can filter the url and the title as listed in [3] and access to the online webcam devices without any access control. As an example;

```
intitle:"Live View / - AXIS" | inurl:view/view.shtml
```

3.4 Secret Data

Secret keys, private keys, encrypted messages compose of the secret data which is expected to be accessible *only* to its owner.

Secret Keys

Normally the secret keys are generated as session keys and destroyed after the session is closed. They are not permanently stored on the disks. But there are certain applications like Kerberos [5] that still need to store a secret key for each principal. The following query searches for dumped Kerberos key databases.

```
"index of" slave_datatrans OR from_master
```

Private Keys

The following search reveals private keys that must be normally kept private.

```
"BEGIN (DSA|RSA)" ext:key
```

Gnupg [1] encodes the private key in secring.gpg files. The following search reveals secring.gpg files.

```
"index of" "secring.gpg"
```

Encrypted Messages

The encrypted files with Gnupg have the extension *gpg*. Signed and public key files have also this extension. The following query searches for files with gpg extension and eliminates non-relevant signed and public key files.

```
-"public|pubring|pubkey|signature|pgp|and|or|release" ext:gpg
```

Mostly the encryption applications use the extension *enc* for the encrypted files. This query searches for the files with the extension enc.

```
-intext:"and" ext:enc
```

In XML security, the encrypted parts of messages are encoded under *CipherValue* tag.

```
ciphervalue ext:xml
```

4 Countermeasures

Google hacking can be very harmful against user privacy and therefore the required security countermeasures should be taken against it. The protection means can be grouped as user-self protection and system-wide protection.

As the name implies, user-self protection requires the users to safeguard themselves against the possible threats. If we enumerate some points which the users should take care of:

- Do not make any sensitive data like documents containing your address, phone numbers, backup directories and files, secret data like passwords, private emails, etc. online accessible to the public.
- Provide only required amount of personal information for the Wiki-similiar management systems.
- Instead of using a single username over the internet, try to hold more pseudonyms which make linkability of user actions through a single username more difficult.
- Considering the forum postings and group mails, try to stay anonymous for certain email contents. Do not mention any company or organization name inside the postings if not required.
- Do not let private media get shared over Web2.0 services.
- Enable authentication techniques for your installed online devices like webcams, printers, etc.

As an administrator, you should focus on system-wide protection for the privacy of their users. The first method you can use is automatic scan tools [4, 7, 8] that search possible Google threats and test privacy risks within your system. The tools mostly use the hack database [3] when they do scan. Another method is integration of robots.txt (robots exclusion standard) [6] files into your system. Web crawlers (*hopefully*) respect the directives specified in robots.txt. Providing this, you can prevent the crawlers from indexing your sensitive files and directories. In addition to this method, you should never put database backups that contain usernames and passwords accessible over your system. The most advanced but also complicated method is installing

and managing Google honeypots [2] in your system and trying to figure out the behavior of attackers before they attack your *real* system.

5 Future Work

The users can be equipped with a penetration testing tool that would search automatically for the possible privacy threats and report its results. Providing this, the users can be aware of the privacy risks which threaten them. We are currently implementing such a tool which will search Google mainly for the privacy risks mentioned in this paper for a specific person and a specific host. Besides, the tool will have a support of finding cryptographic secrets as explained in [9] in details.

References

1. The gnu privacy guard. http://www.gnupg.org.
2. Google Hack Honeypot Project. http://ghh.sourceforge.net.
3. Google Hacking Database. http://johnny.ihackstuff.com/index.php?module=prodreviews.
4. Goolink- Security Scanner. www.ghacks.net/2005/11/23/goolink-scanner-beta-preview/.
5. Kerberos: The network authentication protocol. http://web.mit.edu/Kerberos/.
6. Robots exclusion standard. http://en.wikipedia.org/wiki/Robots.txt.
7. SiteDigger v2.0 - Information Gathering Tool. http://www.foundstone.com/index.htm?subnav=resources/navigation.htm&subcontent=/resources/proddesc/sitedigger.htm.
8. Johnny Long. Gooscan Google Security Scanner. http://johnny.ihackstuff.com/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=33.
9. Emin Islam Tatli. Google reveals Cryptographic Secrets. Technical Report of 1. Crypto Weekend, Kloster Bronbach, Germany, July 2006.