

Privacy in Danger: Let's google Your Privacy

Emin Islam Tatli

Abstract Protection of personal data is a requirement from both ethical and legal perspectives. In the Internet, search engines facilitate our lives by finding any searched information within a single-click time. On the other hand, they threaten our privacy by revealing our personal data to others. In this paper, we give concrete examples of Google personal data exploits against user privacy, discuss the countermeasures to protect our privacy and introduce a penetration testing tool called *TrackingDog* checking and reporting privacy exploits over Google.

Key words: user privacy, Google hacking, privacy exploits, privacy enhancing tools

1 Motivation

Privacy is both an ethical and legal requirement for the Internet users. Protection of personal data against unauthorized access and exploits is inevitable for the users. Different legacy systems in many countries define strict laws to prevent illegal use of personal data [3, 2, 14, 13, 15]. Unlike legacy systems, safeguarding user privacy is not an easy task from the technical point of view. The users share their personal data with third parties, but they can not control whether their data is used for other purposes or forwarded to others. They need to trust the data receivers. P3P (Platform for Privacy Preferences) [12] and Appel (A P3P Preference Exchange Language)[1] projects are two attempts in this direction to build a trust relation between data owners and data receivers.

University of Weimar, Faculty of Media, Chair of Media Security
e-mail: emin-islam.tatli@medien.uni-weimar.de

Web crawlers threaten personal privacy by indexing more and more private data for unauthorized access. The biggest threats can result from the “indexing anything” features of the Internet search engines like Google, Yahoo, Lycos, etc. Especially, Google with its huge index size threatens our privacy. Today, a special research area called *Google Hacking* that focuses on finding vulnerable servers, applications, various online devices, files containing username-password pairs, login forms, etc. exists. We, as the individual users, need to take user-centric countermeasures in order to protect our privacy. In this paper, we illustrate real life examples of privacy exploits via Google hacking, discuss its social aspects and the countermeasures for Google hacking and illustrate our penetration testing tool *TrackingDog* for the user-centric privacy control.

The paper is organized as follows: Section 2 focuses on the real life examples of privacy exploits from Google hacking area. Section 3 explains the possible security countermeasures for Google privacy hacking and introduces our privacy penetration testing tool *TrackingDog*. Finally, Section 4 discusses the privacy policy of Google and proposes some enhancements for better privacy management.

2 A Case Study: Google Hacking for Privacy

Google is the most popular web search engine in the Internet. It indexes any information from web servers thanks to its hardworking web crawlers. But sensitive 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 the privacy risks, there might exist other security threats that can be revealed by Google. There exists even an online database [6] which contains 1423 different Google hacking search queries by November 2007.

2.1 Google Advanced Search Parameters

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

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 apply 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 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*.

2.2 Privacy Searches

Google can be queried for revealing sensitive personal data by using its advanced search parameters. We have grouped private data searches into four different groups according to the privacy level. These are *identification* data, *sensitive* data, *confidential* data and *secret* data searches.

2.2.1 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 the Internet, etc. are the typical examples of the 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 for the web pages and documents which contain keywords like name, surname, address, phone numbers, birthdate, email, etc., optionally for a certain person or within certain document types.

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

Twiki¹ is a wiki-based web application that is commonly used for project management. Inside TWiki, user data like name, address, phone numbers, web pages, location, emails, etc. are stored. If the required authentication techniques are not enforced, unauthorized people can also access this data.

¹ Twiki: <http://twiki.org>

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

In addition to Google search, other search engines with the “people-find” capability can also be very helpful for gaining the identification data. Yahoo’s People Search², Lycos’s WhoWhere People Search³ or eMailman’s People Search⁴ connecting public ldap servers are examples of such services. Similarly, the Firefox plug-in “People Search and Public Record Toolbar”⁵ gives you many facilities to search for the identification data.

Curriculum Vitae

You can search for the keyword CV (curriculum vitae) that can contain the identification data. This search can be extended by searching CV in different languages. For example, Lebenslauf can be used within the search query as the german translation for CV.

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

Login Names

Webalizer application⁶ collects statistical information of web sites about their visitor activities. The most commonly used login names are also stored by Webalizer.

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

2.2.2 Sensitive Data

The sensitive data means that the data which is normally public but its reveal may disturb its owner under certain cases. The examples are postings sent to forums, emails sent to mailing lists, sensitive directories and Web2.0-based social networking applications.

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 People Search and Public Record Toolbar, <https://addons.mozilla.org/en-US/firefox/addon/3167>

⁶ Webalizer: <http://www.mrunix.net/webalizer/>

Forum Postings, Mailinglists

PhpBB⁷ 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"
```

Mailman⁸ 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-based Applications

The next generation Internet Web2.0 introduces more privacy risks. People share more personal data with others within Web2.0-based social networking and blogging applications. The following searches are based on the favorite Web2.0 services like Yahoo's Image Sharing⁹, Google's Blogger¹⁰, Google's Video Sharing¹¹ and Facebook¹². 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 OR site:youtube.com  
"thomas fischer" site:facebook.com
```

⁷ PhpBB Forum: <http://www.phpbb.com>

⁸ Mailman List Manager: <http://www.gnu.org/software/mailman/>

⁹ Yahoo Image Sharing: <http://www.flickr.com>

¹⁰ Google's Blogger: <http://www.blogspot.com>

¹¹ Google Video Sharing: <http://www.youtube.com>

¹² Facebook-Social Networking: <http://www.facebook.com>

2.2.3 Confidential Data

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

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 personal emails within single files like incoming messages as 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 [9] 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 web cameras come along with their software for the remote management over the Internet. Based on the type of the webcam, you can filter the url and the title as listed in [6] and access to the online webcam devices without any access control. As an example;

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

2.2.4 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 [8] that still need to store a secret key for each principal. The following query searches for the dumped Kerberos key databases.

```
"index of" slave.datatrans OR from_master
```

Private Keys

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

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

Gnupg [5] 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
```

The encryption applications mostly 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
```

3 Countermeasures

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

As its 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-similar management systems.
- Instead of using a single username over the Internet, try to have 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 social networking and blogging services.

As an administrator, you should focus on system-wide protection for the privacy of the users as well. The first method you can enforce is using automatic scan tools [7, 10, 17, 18] that search possible Google threats and test privacy risks within your system. The tools mostly use the hack database [6] when they do the scans. Another method is integration of robots.txt (robots exclusion standard) [9] 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 [16] in your system and trying to figure out the behavior of attackers before they attack your *real* system.

3.1 TrackingDog - A Privacy Tool against Google Hacking

To help the users to protect their privacy, new privacy enhancing tools are needed. For example, 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 have already implemented such a tool namely *TrackingDog* [17] which searches Google mainly for the privacy exploits mentioned in this paper for a given person and/or a given host. Besides, the tool has the support of finding cryptographic secrets as explained in [19] in details. *TrackingDog* helps the individuals to detect if any of their confidential data have become public over the Internet via Google. It supports both English and German language-specific queries and enables the users to edit raw search queries. Figure 3.1 illustrates the main GUI of TrackingDog.

4 Discussion

Considering the privacy exploits explained in the previous sections, one can ask himself if such exploits are also misused by Google itself to profile people and track their activities. Even though Google replies this question as *no* and claims to respect our privacy, we can not be sure about this dilemma.

On the other hand, we see some good approaches to privacy by Google. Lately, they have declared that they would take steps to further improve

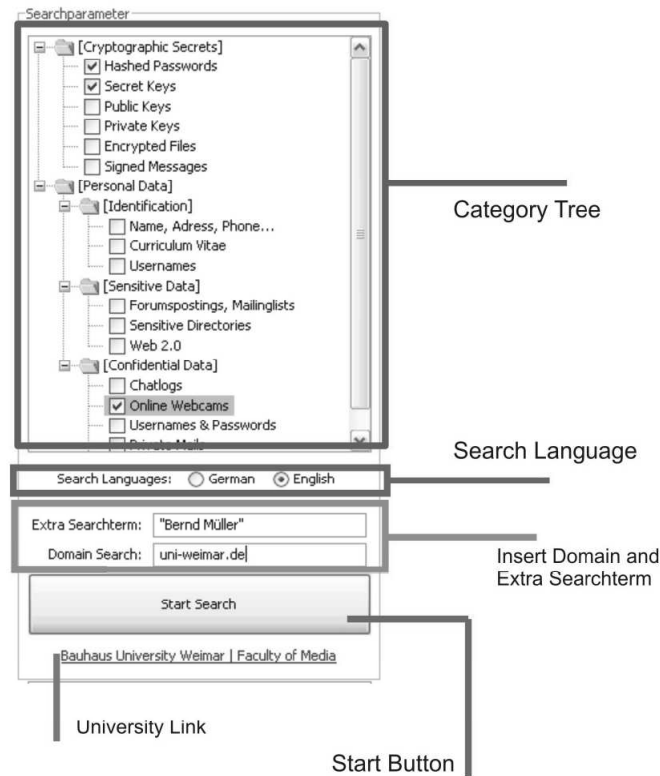


Fig. 1 TrackingDog Main GUI

privacy. By searching in Google, your query, your IP and cookie details are stored on the Google servers and that information can identify you uniquely. But now Google has decided to anonymise this collected data within a 18-24 month period [11]. You can even apply other means to remove your cookies from Google servers as explained in [4].

We believe, Google can do more for our privacy. The privacy exploits mentioned in this paper should be taken into consideration by Google. The personal data should not be collected by the Google crawlers. Internet users are careless and easily make their personal data public unintentionally. This should not be misused by Google. While we hope more respect to our privacy from Google, we also need to have the users get equipped with the powerful user-centric privacy enhancing tools like TrackingDog to get to know the threats and protect themselves.

References

1. A P3P Preference Exchange Language (Appel).
<http://www.w3.org/TR/P3P-preferences/>.
2. EU Directives 2002/58/EC.
http://www.dataprotection.ie/documents/legal/directive2002_58.pdf.
3. EU Directives 95/46/EC.
http://www.cdt.org/privacy/eudirective/EU_Directive_.html.
4. Five Ways to Delete Your Google Cookie. <http://googlewatch.eweek.com/content/five-ways-to-delete-your-google-cookie.html>.
5. The gnu privacy guard. <http://www.gnupg.org>.
6. Google Hacking Database. <http://johnny.ihackstuff.com/index.php?module=prodreviews>.
7. Goolink- Security Scanner.
www.ghacks.net/2005/11/23/goolink-scanner-beta-preview/.
8. Kerberos: The network authentication protocol. <http://web.mit.edu/Kerberos/>.
9. Robots exclusion standard. <http://en.wikipedia.org/wiki/Robots.txt>.
10. SiteDigger v2.0 - Information Gathering Tool.
<http://www.foundstone.com/index.htm?subnav=resources/navigation.htm&subcontent=/resources/proddesc/sitedigger.htm>.
11. Taking steps to further improve our privacy practices.
<http://googleblog.blogspot.com/2007/03/taking-steps-to-further-improve-our.html>.
12. The Platform for Privacy Preferences. <http://www.w3.org/2006/07/privacy-ws/>.
13. Bundesdatenschutzgesetz (BDSG), Germany.
<https://www.datenschutzzentrum.de/material/recht/bdsg.htm>, 1978.
14. Data Protection Act, UK.
<http://www.opsi.gov.uk/acts/acts1998/19980029.htm>, 1998.
15. Bundesgesetz ber den Schutz personenbezogener Daten(Datenschutzgesetz 2000 - DSG 2000), Austria. <http://www.dsk.gv.at/dsg2000d.htm>, 2000.
16. Google Hack HoneyPot Project. <http://ghh.sourceforge.net>, 2007.
17. Martin Kessler. Bachelorarbeit: Implementation of a penetration testing tool for searching cryptographic secrets and personal secrets with Google. Bauhaus Universitaet Weimar, Medien Fakultaet, October 2007.
18. Johnny Long. Gooscan Google Security Scanner.
<http://johnny.ihackstuff.com/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=33>.
19. Emin Islam Tatli. Google reveals Cryptographic Secrets. Technical Report of 1. Crypto Weekend, Kloster Bronbach, Germany, July 2006.