Get to Know Google… Because They Know You

Ethics and Law on the Electronic Frontier, 6.805

By
Stefanie Alki Delichatsios and Temitope Sonuyi

December 14, 2005

Table of Contents

1.0 Introduction
2.0 Personal Data Collection by Google Services
    2.1 Google's Search Engine
        2.1.1 How Google Search Works
        2.1.2 Information Google Records with Google Search
            2.1.2.1 Cookies
            2.1.2.2 IP Address
            2.1.2.3 Browser Configuration, Date and Time, Search Content
            2.1.2.4 Links
    2.2 AdSense
        2.2.1 How AdSense Works
        2.2.2 Information Google Records with AdSense
    2.3 Gmail
        2.3.1 How Gmail Works
        2.3.2 Information Google Retrieves from Gmail
            2.3.2.1 Gmail Registration
            2.3.2.2 Email Content
            2.3.2.3 Server Logs
    2.4 Synthesizing Collected Information
        2.4.1 Scenario of Data Synthesis
            2.4.1.1 Bob Smith's Google Activity
            2.4.1.2 Data Gathered
            2.4.1.3 Separate and Synthesized Data Views
            2.4.1.4 Hypothetical Google profile of "Bob Smith"
            2.4.1.5 Implications of Google Profiling
3.0 Information Processing and Dissemination
    3.1 Google Activity
        3.1.1 The Safe Harbor Program
        3.1.2 Personal Information
        3.1.3 Sensitive Information
        3.1.4 Aggregated non-Personal Information
        3.1.5 Information Processing and Sharing
        3.1.6 Google's Compliance with the Safe Harbor Program
    3.2 Information Released Outside of Google
        3.2.1 Government Officials Requesting Information
        3.2.2 Potential Google Acquirers
4.0 Survey
    4.1 Survey Results
    4.2 Discussion of Survey Results
5.0 Google-Anon
    5.1 Project Goals
    5.2 Project Description
    5.3 Project Challenges and Issues
        5.3.1 Making Users Anonymous
        5.3.2 Getting Past Google's Cleverness
6.0 Conclusions

7.0 References

1.0 Introduction

Google, through its numerous services and popularity, accesses far more information about people than they realize. Though Google explicitly expresses its concern for protecting the vast amount of private user information it collects, that information is nonetheless susceptible to fall into the hands of a) government officials seeking information through warrants, court orders and subpoenas, and b) potential Google acquirers. Google users must be made aware of the personal user information Google collects and what does and can happen to that information. While some users may not be bothered by Google's data collection, others might feel extremely violated and may choose to behave differently when using Google's services. In either case, Google users who have been made aware of these privacy issues and presented with anonymous alternatives, *will* gravitate towards using these alternatives.

In Section 2 of this paper, we examine what exact personal user information Google collects with its three most widespread services, Google Search, AdSense, and Gmail, and how this information can combine to create large identifying profiles about its users. In Section 3, we explore Google's explanation for handling this personal information and how it adheres to the Safe Harbor Program. We also present ways in which this private information can potentially escape the confines of Google's private servers, specifically through government subpoenas and corporate acquisitions. In Section 4, we discuss the results of a survey we submitted to 60 internet users about their understanding of Google's privacy issues and their interest in anonymizing themselves from Google.

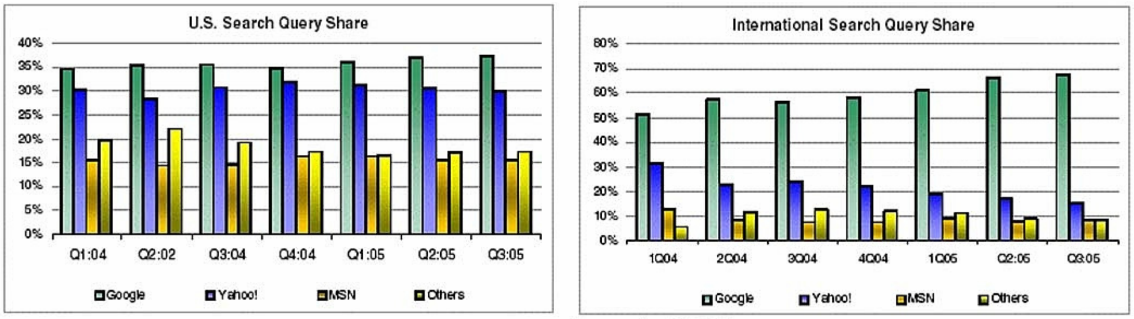
Finally, in an effort to help Google users be more conscious about their Google searches, we present a service called Google-Anon in Section 5, allowing users to search Google anonymously and compare the differences of search results based on different IP addresses and the presence of cookies. With Google-Anon, a user has the option to either a) search Google anonymously through a routing network as a direct substitute to searching with Google regularly or b) view a comparison of differences, for example, the number of results returned or the type and language of the ads presented, between a search made through an anonymized network with no cookies and a search made using a user's specific IP address and Google cookies.

2.0 Personal Data Collection by Google Services

Google, Inc., formed in 1998 as a simple search engine responding to 10,000 queries per day, has transformed into a multinational corporate leader providing over 30 widely used services with a search engine that now answers over 200 million queries per day [1][2]. By combining information from its different services through Google cookies and other logging information, Google has the ability to create huge dossiers of personal information about its individual users. Though some of Google's smaller services, such as Google Desktop[1] and Google Toolbar[2] are more obviously penetrating, we choose to examine Google's three most extensive yet unassuming services, Google Search, AdSense and Gmail to demonstrate how the information Google collects from these three services can combine to produce an alarmingly large "profiles" of its individual users.

2.1 Google's Search Engine

Google's search engine stemmed from a Stanford PhD project, "BackRub" in 1996, and 9 years later, is the leading internet search engine over others like Yahoo and MSN, answering over 35 percent of U.S. internet searches and over 65 percent of international internet searches. [5]:



(a) **(b)**

Figure 2.1 Google dominates in both the (a) U.S. search market and the (b) international search market .

The success of Google's search engine can be attributed to its uncluttered interface, its unobtrusive advertisements, and most importantly, its trademarked ranking system, PageRank^{TM}.

2.1.1 How Google Search Works

Google crawls the web and currently indexes over 9 billion items [6]. Like other search engines, Google organizes web pages by their content-- the frequency of words on a page, the position of words on the page, and the font size and capitalization of words [7]. When a user makes a request on Google, Google uses content information to match the request. Google then combines a document's content information with its PageRank to determine the ordering of the sites returned to the user. The PageRank of website A, is determined by the number of other websites linked to website A, and the quality and PageRank of those linked websites [8].

[1] Google Desktop is a desktop search application giving Google access to information and files on a user's hard drive. It also provides a "sidebar" for the user to easily view customized personal information, such as weather and news, from the web [3].

[2] Google Toolbar is a internet toolbar service with advanced features such as WordTranslator and SpellCheck. When a user chooses to enable "advanced features", all of the user's internet activity is logged by Google [4].

2.1.2 Information Google Records with Google Search

It is well-known that behind its simple interface, the Google search engine performs complicated algorithms on billions of existing websites to maximize the quality of a user's search. However, what people do not realize is that the engine *also* collects and processes massive amounts of information about the individual searcher.

Google records a "server log" *every* time a user makes a query with Google's search engine. The server log includes the user's cookies, IP address, browser type, browser language, data and time of request, and the search content [9].

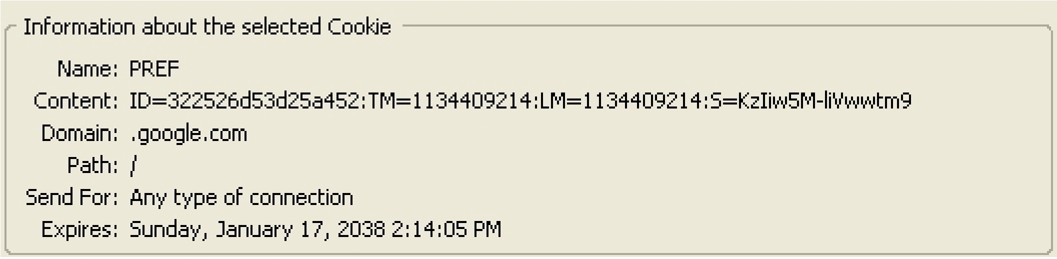
A typical server log where the search is for "dictionary" may look like this:

```
18.127.42.66 – 5/Dec/2005 9:20:46 –
http://www.google.com/search?q= dictionary – Firefox 1.0.7; Windows NT 5.1 –
740674ce123969
```

2.1.2.1 Cookies

In the example server log, "740674ce123969" refers to the user's cookie. A cookie is a unique ID placed on a user's hard disk. Every time a user does a Google search, Google places a cookie on the user's machine if it does not already have one. If the user already has a Google cookie on his or her machine, Google can read and record the cookie [10]. Google's cookies expire over thirty years from their initial formation [10]. While computer users have the option to erase their cookies, most do not, allowing Google to link a person's cookie with other information it collects about a user as long as that user has the same computer.

To verify Google's cookie management, I erased all my internet cookies and went to Google.com. Without making a search, Google placed a cookie on my machine with the following information:



Information about the selected Cookie

```
     Name: PREF
  Content: ID=322526d53d25a452:TM=1134409214:LM=1134409214:S=KzIiw5M-liVwwtm9
   Domain: .google.com
     Path: /
 Send For: Any type of connection
  Expires: Sunday, January 17, 2038 2:14:05 PM
```

Figure 2.2 Google cookie placed on user's machine after visiting Google.com. Notice how the cookie does not expire until 2038.

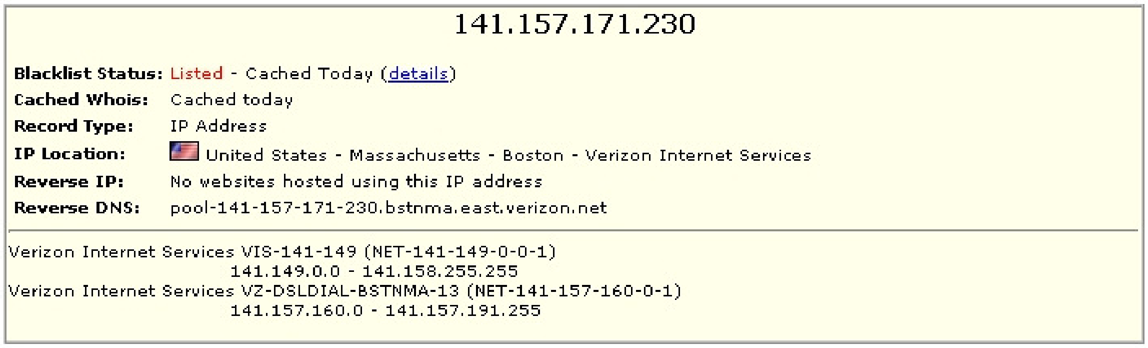
Whenever I clear my cookies and then visit a website anywhere within the google.com domain, a similarly formatted cookie gets placed on my machine.

2.1.2.2 IP Address

The number “18.127.42.66” is an internet user’s internet protocol (IP) address, a 32-bit unique number that a computer uses to identify and communicate with other computers on an IP network, i.e. the internet [11]. A user’s IP address is assigned by his/her Internet Service Provider (ISP). The Internet Assigned Numbers Authority (IANA) assigns local registrations of IP addresses to five Regional Internet Registries (RIRS)[3] which are then responsible for allocating IP addresses to ISPs within their region [12] [13]. A user’s IP address may or may not change each time the user connects to the internet, but in either case, the IP address reveals location-specific information about the user.

Many computer networks today are connected to the internet through Network Address Translation (NAT) [14]. With the increase of internet users, especially within home and business networks and the way in which sections of the IP address spectrum are blocked and reserved for specific purposes, there are simply not enough available IP addresses. As a result, NAT allows for many computers on an internal network to connect to the internet by sharing a single IP address through a router [14]. Again, though a user’s computer may share an IP address with many others, that IP address is still very telling about the user’s geography.

With the knowledge of a user’s IP address, anyone can simply discover location information about that user. Many websites on the internet exist providing a “who is” service, allowing a user to retrieve information about a particular IP address. For instance, if I am surfing the internet from my apartment in Cambridge, I can go to whatismyip.com to determine my computer’s IP address. When I do that, I discover my IP address is: 141.157.171.230 [15]. I can then go to “Whois Source”, http://www.whois.sc/, to obtain information about my IP [16]:



```
                         141.157.171.230

Blacklist Status: Listed - Cached Today (details)
Cached Whois:     Cached today
Record Type:      IP Address
IP Location:      United States - Massachusetts - Boston - Verizon Internet Services
Reverse IP:       No websites hosted using this IP address
Reverse DNS:      pool-141-157-171-230.bstnma.east.verizon.net

Verizon Internet Services VIS-141-149 (NET-141-149-0-0-1)
                          141.149.0.0 - 141.158.255.255
Verizon Internet Services VZ-DSLDIAL-BSTNMA-13 (NET-141-157-160-0-1)
                          141.157.160.0 - 141.157.191.255
```

Figure 2.3 Doing an IP “lookup” on *Whois Source* reveals user’s location and ISP.

Whois Source provides a detailed description of my IP address, allowing a user to quickly and easily detect that I am located in the Boston area and that I use Verizon as my internet service provider. Since Google records a user’s IP address as part of its server log, it too can trace the geographic location of an individual user.

[3] The five RIRS are AfriNIC (African Network Information Centre, APNIC (Asia Pacific Network Information Centre), ARIN (American Registry for Internet Numbers), LACNIC (Latin American and Caribbean IP address Region Registry), and RIPE NCC (Reseaux IP Europeens) [12].

2.1.2.3 Browser Configuration, Date and Time, and Search Content

Finally, "Firefox 1.0.7; Windows NT 5.1" from the server log refers to the user's browser and operating system, "5/Dec/2005 9:20:46" is the date and time of the search, and http://www.google.com/search?q= dictionary" is the requested URL, with the query "dictionary" included.

2.1.2.4 Information about Links

On Google's "Search Results" pages, Google records the fact that a user clicked on a link and that link's URL in order to "determine how often users are satisfied with the first result of a query and how often they proceed to later results" [9]. Essentially, Google tracks "where" a user goes after he or she leaves Google's Results Pages.

2.2 AdSense

The AdSense program, a Google service developed in 2003, allows a website to host contextualized advertisements, "AdWords", and generate revenue on a cost-per-click (CPC) basis [1].

2.2.1 How Adsense Works

A website using AdSense integrates a piece of Javascript code into the site's HTML which allows Google to control the type, placement, and number of advertisements on that particular website [17]. Google uses the content of the website to select appropriate advertisements for the website [17]. Additionally Google factors the language of the site and the location of the visitors to enhance the relevancy of the advertisements [18].

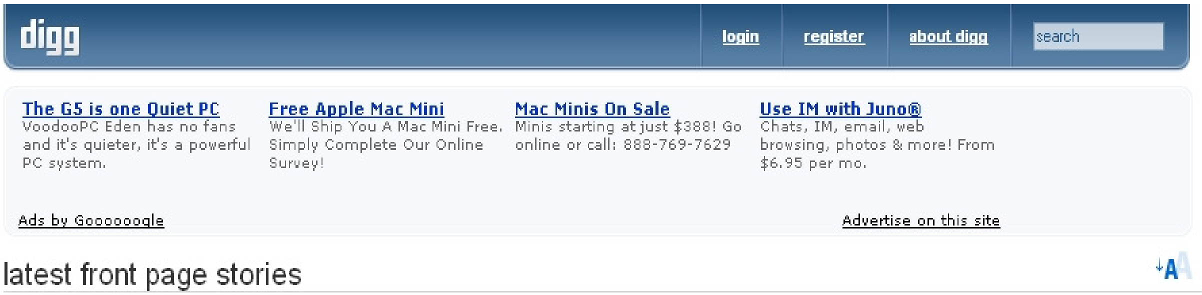
For example, "digg.com" is a technology news website whose source code reveals its use of AdSense:

```
<div id="google-broad"><div><script type="text/javascript"><!--
google_ad_client = "pub-7489042062340760";
google_ad_width = 728;
google_ad_height = 90;
google_ad_format = "728x90_as";
google_ad_type = "text";
google_ad_channel ="4567327683";
google_color_border = "F7F8FB";
google_color_bg = "F7F8FB";
google_color_link = "0033CC";
google_color_url = "0066CC";
google_color_text = "666666";
//--></script>
<script type="text/javascript"
  src="http://pagead2.googlesyndication.com/pagead/show_ads.js">
</script>
```

Google programs the type, language, and colors of the advertisements in the javascript file, http://pagead2.googlesyndication.com/pagead/show_ads.js. A piece of the code is shown below:

```
google_append_url('dt', date.getTime());
  google_append_url('hl', w.google_language);
  if (w.google_country) {
    google_append_url('gl', w.google_country);
  } else {
    google_append_url('gl', w.google_gl);
  }
  google_append_url('gr', w.google_region);
  google_append_url_esc('gcs', w.google_city);
  google_append_url_esc('hints', w.google_hints);
  google_append_url('adsafe', w.google_safe);
  google_append_url('oe', w.google_encoding);
  google_append_url('lmt', w.google_last_modified_time);
  google_append_url_esc('alternate_ad_url',
                               w.google_alternate_ad_url);
  google_append_url('alt_color', w.google_alternate_color);
  google_append_url("skip", w.google_skip);
```

Google determines the country, region, and city of the user and in doing so, chooses appropriate advertisements for the site. When I visit digg.com from my apartment in Cambridge, MA, the advertisements appear like this:



The G5 is one Quiet PC
VoodooPC Eden has no fans and it's quieter, it's a powerful PC system.

Free Apple Mac Mini
We'll Ship You A Mac Mini Free. Simply Complete Our Online Survey!

Mac Minis On Sale
Minis starting at just $388! Go online or call: 888-769-7629

Use IM with Juno®
Chats, IM, email, web browsing, photos & more! From $6.95 per mo.

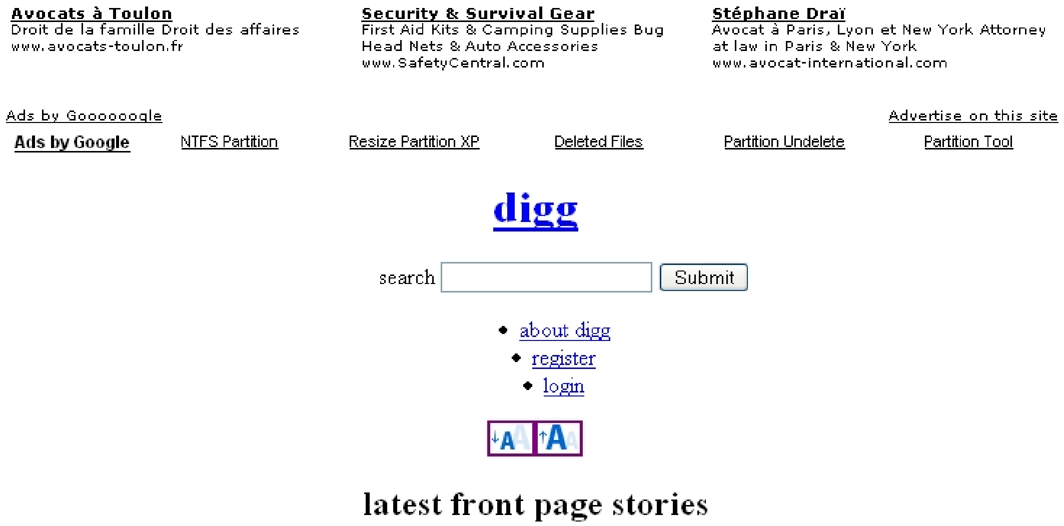
Ads by Goooooogle

Advertise on this site

latest front page stories

Figure 2.4 When user with American IP address visits digg.com, Google Ads are in English.

When a user in France visits digg.com, the advertisements appear like so:



Avocats à Toulon
Droit de la famille Droit des affaires
www.avocats-toulon.fr

Security & Survival Gear
First Aid Kits & Camping Supplies Bug Head Nets & Auto Accessories
www.SafetyCentral.com

Stéphane Draï
Avocat à Paris, Lyon et New York Attorney at law in Paris & New York
www.avocat-international.com

Ads by Goooooogle | Advertise on this site

Ads by Google | NTFS Partition | Resize Partition XP | Deleted Files | Partition Undelete | Partition Tool

digg

search [ ] Submit

- about digg
- register
- login

↓A ↑A

latest front page stories

Figure 2.5 When user with French IP address visits digg.com, Google Ads are in French.

Google combines information from the content of the page with information from the user's IP address to target the ads to the individual user.

AdSense users generate revenue either on a CPC[4] (cost-per-click) or CPM[5] (cost per thousand impressions). Adsense users can also choose to host a Google search bar on their websites, allowing the the site's users to search Google directly from the website. The AdSense user profits from the search bar by the advertisements shown on the first results page of the query.

2.2.2 Information Google Retrieves from AdSense

Each time a user visits a website with AdSense, Google records a server log similar to the log recorded with Google Search. Instead of tracking http://www.google.com/search?q=dictionary, the log simply records the URL of the visited site.

Google also tracks each time a user clicks on one of the advertisements as part of the CPC paying method.

2.3 Gmail

[4] CPC (cost-per-click) refers to the amount paid by the AdWords user every time someone clicks on his/her advertisement. With AdWords, an advertiser choses a maximum CPC from $.01-$100 [19]
[5] CPM (cost per thousand impressions) refers to the amount paid by the AdWords user for each 1000 of his/her ads shown [19].

Gmail, released in April 2004, is Google's free search-based webmail service supplying its users with over 2.5 gigabytes of storage [20].

2.3.1 How Gmail Works

Gmail works like any webmail service, but differs in its its powerful search engine and its focus on the virtually unlimited storage it provides. Google encourages its Gmail users to perform easy and quick searches instead of creating folders and filing messages [20].

2.3.2 Information Google Retrieves from Gmail

With Gmail, Google retrieves personal user information from account registration and email content. Additionally, as with Google Search and AdSense, Google creates a server log every time a user visits the Gmail website, linking log information such as the user's cookies and IP address with the user's personal Gmail information.

2.3.2.1 Gmail Registration

Google requires an invite from a current Gmail user in order for a new user to create a Gmail account. If the new user does not have an invite, the user may request an invite through a mobile text message. If the user chooses to receive an invite through his mobile phone, Google records the user's mobile phone number [22].

As part of the user account registration, Google requests the user's first and last names, and a secondary email address of the user.

2.3.2.2 Email Content

Google scans email content, as all email providers do, to provide spam filtering, virus detection, search and other services [22]. Gmail also uses email content to provide target-based advertisements.

Gmail maintains several backup copies of users' emails to recover messages and restore accounts in case of system failure.

If a user deletes an email or terminates his Gmail account, Google reflects these actions in the user's account view. However, "residual copies of deleted messages and accounts may take up to 60 days to be deleted from [their] active servers and **may remain in [their] offline backup systems**". An email that a user intended and expected to be erased may in reality remain on Google's servers forever [22].

2.3.2.3 Server Logs

Google records a server log for Gmail activity just as it does with Google Search and AdSense. In addition to the basic log information (i.e. IP address, date and time, etc), Google also logs account activity, such as storage usage and number of log-ins, and data displayed and clicked on [19].

2.4 Synthesizing Collected Data

Evidently, even with its three most basic and unassuming services, Google tracks every single action made by its users. And though it is unclear what Google does with this information beyond target-based advertisements, Google *can* easily link user activity across different services using a user's cookies, IP address, or Gmail account to create individual user profiles. With Google Search, AdSense, and Gmail alone, Google collects and has the capability to interconnect the following information:

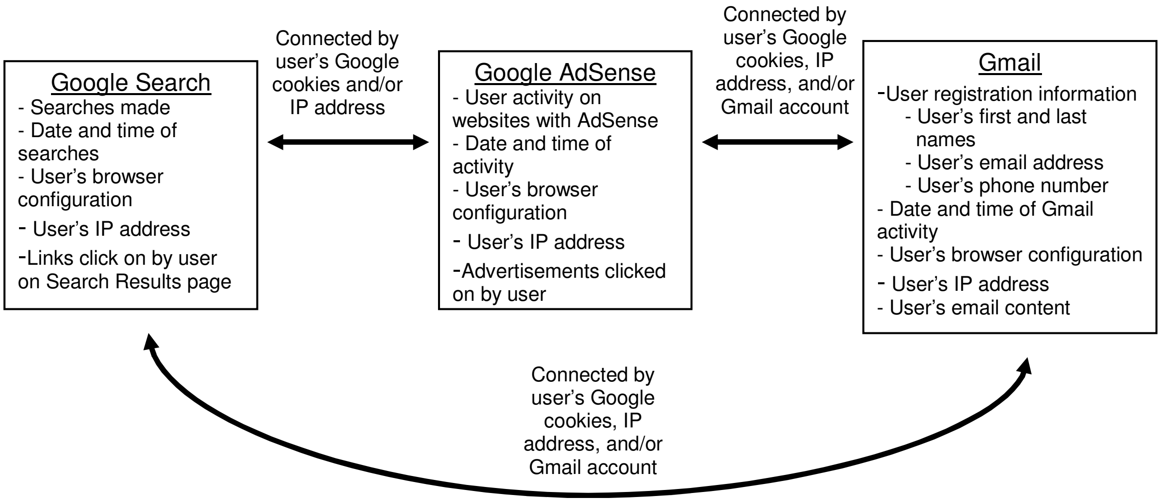

Google Search
- Searches made
- Date and time of searches
- User's browser configuration
- User's IP address
-Links click on by user on Search Results page
Connected by user's Google cookies and/or IP address
Google AdSense
- User activity on websites with AdSense
- Date and time of activity
- User's browser configuration
- User's IP address
-Advertisements clicked on by user
Connected by user's Google cookies, IP address, and/or Gmail account
Gmail
-User registration information
- User's first and last names
- User's email address
- User's phone number
- Date and time of Gmail activity
- User's browser configuration
- User's IP address
- User's email content
Connected by user's Google cookies, IP address, and/or Gmail account

Figure 2.6 Visual representation of collected data from three of Google's services, Google Search, AdSense, and Gmail, and how the information can be interlinked by cookies, IP addresses, and/or Gmail accounts

In essence, Google can interconnect a user's name and email address (from Gmail) and his approximate geographic location (from user's IP address) with particular searches he made (from Google Search) or websites he visited (from AdSense) at specific dates and times.

2.4.1 Scenario of Data Synthesis

To elucidate this idea of data synthesis and its implications, we will present a fictional scenario. The scenario will chronicle a typical user's activity with Google Search, AdSense, and Gmail. We will present plausible data gathered from the user's activity and then provide a synthesized view of the collected data.

2.4.1.1 Bob Smith's Google Activity

- Bob Smith, an MIT student, is sitting in his dorm room one day and goes to Google.com and types in a search for "marijuana coffee"

- On the Search Results page, he clicks on “Amsterdam’s Marijuana Cannabis Coffee Shop Listing”, www.onlinepot.org/amsterdam/amsterdamlist.htm
- Bob then logs into his unique Gmail account which he has set up using his real name and phone number, and reads some mail
- A few days later, Bob goes to his home in San Francisco for Thanksgiving break and once there, uses his laptop to visits his favorite site digg.com, which has GoogleAds on the page
- He then logs into Gmail and checks his mail
- After returning to MIT and continuing to check his email on a regular basis, he decides to erase all his cookies
- Bob then searches for “wedding rings” using Google.com
- Finally Bob again logs into his Gmail account to check mail

2.4.1.2 Data Gathered

- When Bob searches for “marijuana coffee” a cookie is placed on his machine if it doesn’t already exist, and in either case, Google logs his search and the cookie/time/date/ip associated with it
- When Bob clicks on a link from the Google results page, the link he follows is logged, along with the cookie/time/date/ip associated with his click
- When Bob logs into his Gmail account, his activity is logged, along with the cookie that is on his machine at the time. This is done every time Bob logs into his unique Gmail account.
- When Bob visits digg.com, the GoogleAds section of the site gets the Google cookie on Bob’s machine and logs what site he has just visited by associating the page with the cookie. Google also recognizes that Bob is using a different IP address to access the internet

2.4.1.3 Separate and Synthesized Data Views

After Bob searches “marijuana coffee”, he has cookie1 with “marijuana coffee” and other info recorded

Search Terms	Date&Time	IPAddress	CookieID
Marijuana coffee	11/20/2005 09:20:46EST	18.127.42.66	Cookie#1

After Bob follows www.onlinepot.org/amsterdam/amsterdamlist.htm, he has cookie1 with site link and other info recorded

Search Link Followed	Date&Time	IPAddress	CookieID
www.onlinepot.org/amsterdam/amsterdamlist.htm	11/20/2005 09:26:46EST	18.127.42.66	Cookie#1

After Bob goes home to San Francisco and visits digg.com, he has cookie1 and the visited URL recorded

Visited Adsense Site	Date&Time	IPAddress	CookieID
http://www.digg.com	11/24/2005 14:20:46PST	66.127.42.3	Cookie#1

After Bob logs into Gmail account he has his unique Gmail account (name/content), cookie1 and other info recorded

Gmail Account	Date&Time	IPAddress	CookieID	Mail Content
ID#:334 Name: Bob Smith	11/24/2005 22:20:46PST	66.127.42.3	Cookie#1	Mailbox#:334

When Bob returns to MIT, he makes a new Google search with a new cookie, and cookie2 and other info are recorded

Search Terms	Date&Time	IPAddress	CookieID
wedding rings	11/28/2005 05:22:36EST	18.231.4.216	Cookie#2

Bob logs into Gmail again and his Gmail account info and cookie2 with other info are recorded

Gmail Account	Date&Time	IPAddress	CookieID	Mail Content
ID#:334 Name: Bob Smith	11/28/2005 6:20:46EST	18.231.4.216	Cookie#2	Mailbox#:334

2.4.1.4 Hypothetical Google Profile of “Bob Smith”

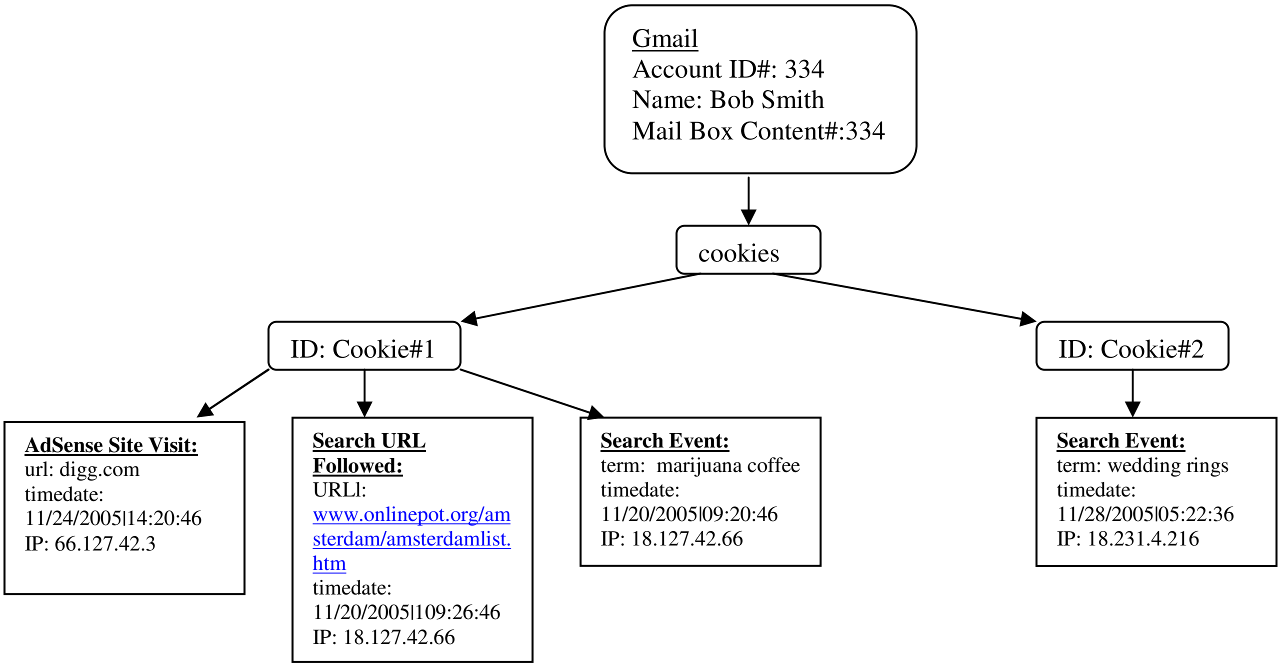


Figure 2.7 Hypothetical Google profile of “Bob Smith” links information about searches Bob has made, sites he has visited, what IP address he uses, and his Gmail activity.

After minimal Google activity on Bob’s part, Google now has a comprehensive profile on a man named “Bob Smith”, including details about his Gmail activity, what searches he’s made and what sites he’s visited when and from what IP address. Midway through his activity, Bob erases his cookies, but nevertheless, when he signs into his Gmail account, he is re-linked to his previous cookie, and all the information from the two cookies are interconnected. With his changing IP address, Google can also trace “Bob Smith”’s geographical movement- logging in from MIT for a couple days, from San Francisco for the next couple of days, and then from MIT again.

Though Google does not explicitly concede to creating such profiles, the privacy policy does state that “we may combine the information you submit under your account with information from other Google services or third parties in order to provide you with a better experience and to improve the quality of our services” [23]. The creation of such a profile is highly realizable.

2.4.1.5 Implications of Google Profiling

When Bob Smith searches “marijuana coffee”, visits digg.com, checks his Gmail account, etc, he is not intending nor is he aware that Google logs and possibly interconnects all of his Google activity. Essentially, Google is recording and possibly synthesizing personal data about Bob Smith that did not exist before.

Google remembers a user's search long after that user has made that search and forgotten it. Google retains emails that a user erases. Some people may not be bothered that Google records and stores all this personal data. However, other users may feel extremely violated by this data collection, regardless of the possibilities of whose hands this information could fall into.

Knowing all this information, bothered users may choose to use Google services differently, either by searching Google anonymously with our proposed anonymizer Google-Anon, choosing not to use Gmail, erasing cookies on a regular basis, or through other methods of preventing Google from collecting personal data. We feel that users should simply be aware that Google collects all this information about its users, and with that knowledge, decide whether or not to change their activity with Google's services.

3.0 Information Processing and Dissemination

Beyond simply the collection of personal information, Google users should be aware of what does and can happen to that information.

3.1 Google Activity

Google's privacy policy, dated October 14, 2005, is explicit and detailed about the different types of information it handles and what sort of analysis it performs on each type of data. Google also states in its policy that it is a registered organization with the U.S. Department of Commerce's Safe Harbor program [23].

3.1.1 The Safe Harbor Program

The U.S. Department of Commerce's Safe Harbor program was developed in 2000 in response to the European Commission's Directive on Data Protection[6] to provide U.S. companies a means to comply with the Directive and avoid facing prosecution by European authorities under Europe's strict privacy laws [24]. Companies registered with the Safe Harbor Program are deemed "adequate" under the European Directive. In order for a company to register with the Safe Harbor Program, it must comply with the seven safe harbor principles: notice, choice, onward transfer, access, security, data integrity, and enforcement [24]. In essence, a company's user must be notified about the purposes for the company's collected personal data, the user must have the opportunity to "opt out" of providing personal information and "opt in" of providing sensitive information, the user must be granted access to any information the company may have about that user, and the user information must be relevant and correct.

Google's privacy policy explains the measures it takes to comply with the Safe Harbor Program. It delves into the types of information it collects and what it does with that information.

[6] The European Commission's Directive on Data Protection, enacted in 1998, prohibited the transfer of personal data to non-European Union nations that did not meet the European "adequacy" standard for privacy expectation [25].

3.1.2 Personal Information

Google describes "personal information" to be information that personally identifies a user, such as a user's name, email address or billing information [9].

3.1.3 Sensitive Information

"Sensitive personal information" refers to a user's confidential medical information, racial or ethnic origins, political or religious beliefs or sexuality that can be connected to the user's personal information [9].

3.1.4 Aggregated Non-Personal Information

"Aggregated non-personal information" is information about a user's Google activity that is collected into groups and does not reference an individually identifiable user [9].

3.1.5 Information Processing and Sharing

Google processes a user's "personal information" to customize content and advertising for the user, to improve Google's services and to develop new services. Google provides personal information to affiliated companies that process information on Google's behalf and are required that they comply with Google's privacy policy. If at any time Google wants to share personal data with companies or persons outside of Google, it will notify its users and provide an opt-out option [23].

Google never processes or shares "sensitive information" without opt-in consent.

Google processes and shares "aggregated non-personal information" with companies and persons outside of Google. See Google Zeitgeist http://www.google.com/press/zeitgeist.html for interesting analysis Google does with aggregated information about user behavior and patterns with the search engine.

3.1.6 Google's compliance with the Safe Harbor Program

Google's privacy policy is essentially written to reflect its registration with the Safe Harbor program with headings such as "Information Security", "Data Integrity", and "Enforcement". Even to the extent that Google creates personal user profiles such as "Bob Smith" from Section 2.4.1.4, Google is in compliance with the Safe Harbor Program and users can somewhat rest assured that their private user information is being carefully managed.

Nonetheless, Google also states that its privacy policy could at any time change yet interestingly, it does not disclose how long it retains the personal user information it collects [23]. If at some point Google decides to change its policy and no longer wishes to be a registered member of the Safe Harbor program, Google would still have all of its collected personal user information, but could choose not to protect that information in the same way it does now.

3.2 Information Released Outside of Google

Though Google is currently dedicated to protecting the vast amount of private personal information it has about its users, that information could easily end up outside of Google's servers, namely through government officials or potential Google acquirers.

3.2.1 Government Officials Requesting Information

In its privacy policy, Google explicitly states that it complies with "valid legal process, such as search warrants, court orders, or subpoenas seeking personal information" [9]. Google may have personal and sensitive information about a user that it protects with the highest level of privacy but at any time, a government official with a warrant could easily come to Google and request that sensitive information about a user and Google would release the information. Furthermore, under Provision 213 of the USA Patriot Act, government officials can request information from Google without notifying the Google user until after the search has happened [26].

In October 2005, Google searches made by an accused murder were brought and used in court by the prosecution. The body of Robert Petrick's wife was found in January 2003 in Falls Lake, North Carolina, and prosecutors discovered Google queries Petrick had made on his computer just prior to his wife's death, including "neck", "snap", "break", and the lake levels and water currents of Falls Lake [27]. Though the searches were found on Petrick's hard drive, not through subpoenaed information from Google, this scenario nonetheless highlights the type of information Google collects and stores and how it is not protected from government intrusion. At the time he made those queries, Petrick did not realize that the information would be stored and released.

3.2.2 Potential Google Acquirers

Google's dedication to protecting its user's privacy is highly respectable but its policy holds only so long as Google is in control of its collected information. In its privacy policy, Google states that in the case of a merger or acquisition, it will "provide notice before personal information is transferred and becomes subject to a different privacy policy" [23]. Google does not and can not guarantee that should a merger or acquisition occur, the personal information it stores will be protected in the same way it is now.

Additionally, if Google is acquired by a company and privacy rights are violated, Google can not be held liable under the Homeland Security Act [28]. The Homeland Security Act protects companies from lawsuits and government prosecution when they turn over information to a new agency.

4.0 Survey

In order to verify how much of the aforementioned information users are actually aware of, we conducted an online survey and asked 60 people the following questions:

1. Is Google your primary search engine? Yes/No. .

2. Have you read Google's Privacy Policy? Yes/No.

3. Are you aware that Google keeps records every search **you** make on **your** machine? Yes/No.

4. Do you have a Gmail account? Yes/No.

5. Are you aware that when you erase an email, Google retains that email on one of their servers? Yes/No.

6. Do you know that government officials can subpoena information Google collects about its users? Yes/No.

7. Google can trace and interconnect
    a) When and what you searched with Google Search
    b) Certain websites you visit that use AdSense
    c) Your name and secondary email address (provided in Gmail registration)
    d) The content of your Gmail email
    e) When and how many times you log into Gmail

*the listed items are pieces of information Google collects from its three most pervasive services- Google Search, AdSense, and Gmail. Feel free to ask more about these services

Knowing this information, do you think you will change your behavior when using Google's services? Yes/No.

8. Would you be interested in an anonymizer that allows you to search Google anonymously? Yes/No.

To see the online version, click on this link:
http://FreeOnlineSurveys.com/rendersurvey.asp?id=134247.

4.1 Survey Results

We distributed the survey by email and instant messenger over a span of two days. Ninety percent of the respondents are students and young professionals ages 20 – 24 and the other ten percent are family members ages 30+.

	Q1		Q2		Q3		Q4	
Yes	98.3%	58	3.3%	2	26.7%	16	71.7%	43
No	1.7%	1	96.7%	58	73.3%	44	28.3%	17
Total Responses		59		60		60		60

	Q5		Q6		Q7		Q8	
Yes	30.5%	18	25.4 %	15	36.7%	22	85%	51
No	69.5%	41	74.6%	44	63.3%	38	15%	9
Total Responses		59		59		60		60

Table 4.1 Summary of survey results, listing the total responses, the number of and percent total "yes" responses, and the number of and percent total "no" responses for each question.

Questions 1 and 4 highlight the pervasiveness of Google's services:

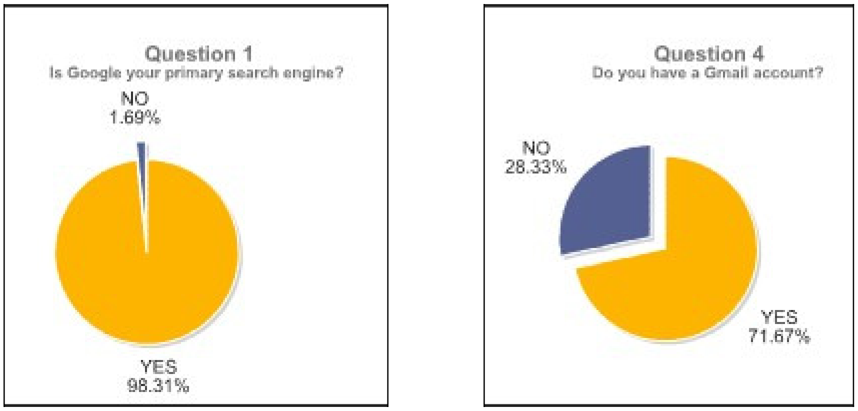


Figure 4.1 Results of Questions 1 and 4 of survey.

Only one person responded saying Google was not his/her primary search engine and over 2/3 of the respondents were registered Gmail users, despite Gmail being a relatively new webmail service.

Questions 2, 3, 5, and 6 demonstrate how little people know about Google's privacy policies:

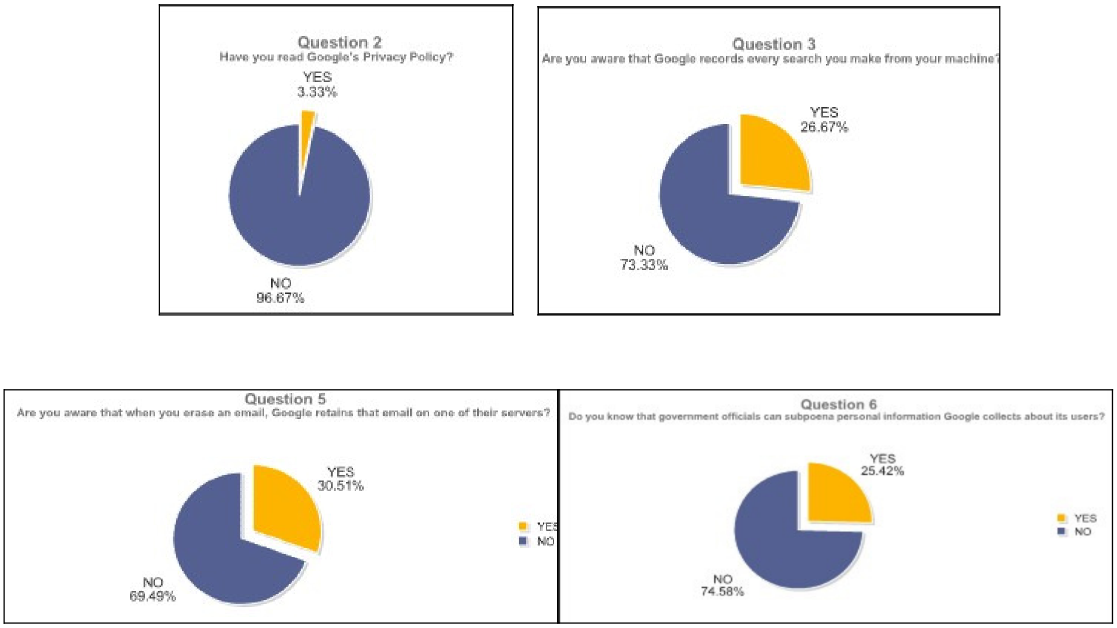


Figure 4.2 Results of Questions 2,3, 5 and 6 of survey.

In Question 7, the majority of people said they would not change their internet behavior when using Google services after realizing what sort of information Google actually collects:

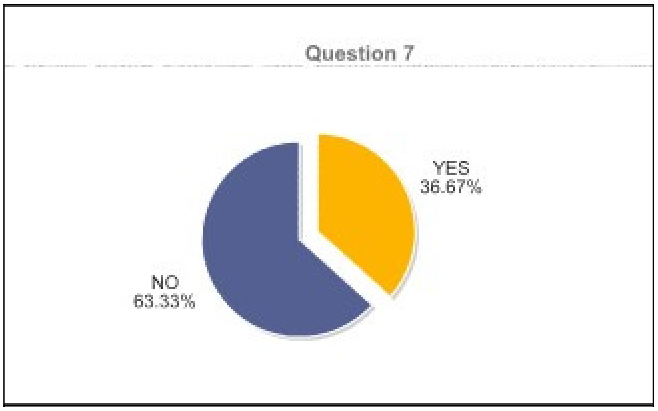


Figure 4.3 Results of Question 7 of survey.

At the same time, most of our survey-takers *were* interested in a service allowing them to search with Google anonymously:

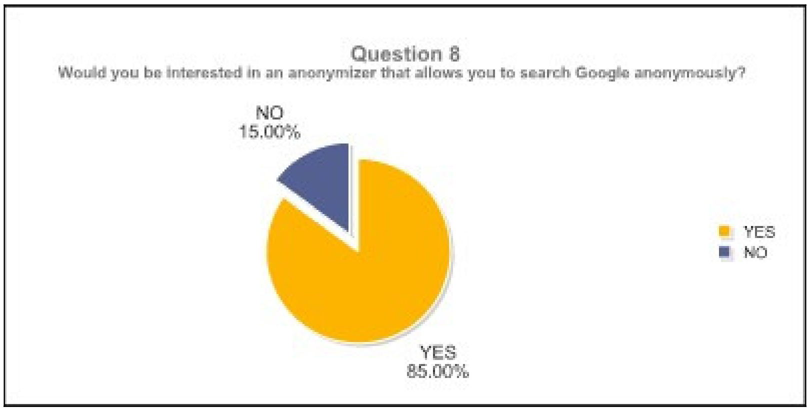


Figure 4.4 Results of Question 8 of survey.

4.2 Discussion of Survey Results

As expected, most of the respondents use Google as their primary search engine and over 2/3 are registered Gmail users. However, it is interesting to note that even among the Gmail users, hardly any of them have read Google's privacy policy. If people actually read Google's privacy policy and responded "yes" to question 2, they would also have responded "yes" to questions 3, 5, and 6. Google thoroughly reveals all relevant privacy information (i.e. questions 3, 5, and 6) in its privacy policy but most people simply have not made that step to read Google's policy.

Though the majority of the survey-takers said they would not change their Google behavior after recognizing Google's personal information collection, over 1/3 still said they would. This statistic highlights the importance of alerting people of Google's data collection. People must be aware of Google's collection of personal user information and with that knowledge, can then decide whether or not to behave differently when using Google's services.

Despite the results of question 7, most of the respondents said they *would* be interested in a Google anonymizer in question 8. This incongruity reveals that question 7 may have been somewhat unclear. If instead, we had posed ways in which a user could change his or her internet behavior like below:

> "Knowing this information, do you think you will change your behavior **(i.e. erase internet cookies on a regular basis, use anonymizers, disable javascripts for pages with Google Ads, etc**.) when using Google's services?",

We expect that the results of question 7 would more closely resemble the results of question 8.

4.3 People's Reactions to the Survey

The most interesting results of the survey were in fact the reactions of survey-takers after completing the survey. One respondent replied in an email saying,

"Your survey is...uh...kinda scary".

Another responded through instant messenger saying,

"Are you trying to make me paranoid? Because it's working".

Yet another responded through instant messenger with,

"Is this true? Can I forward this survey to my friends? I've been to Cuba and that's illegal! And I use my Gmail account all the time- and I wrote a lot of emails about being in Cuba!! I'm so scared!".

It seems that most of the paranoid reactions were in response to questions 3, 5 and 6. This paranoia comes as no surprise simply because personal user information that it is out of the user's control is a new concept to grasp.

Before the advent of the internet, a person's minor actions could not feasibly have been recorded. But with the seemingly unlimited digital memory storage that exists today, entities like Google can easily record and store information about a user that the user has long forgotten. Just as Robert Petrick from Section 3.2.1 did not fathom that Google searches he made would one day be used against him in court, people generally do not like the idea that information about them which they have forgotten or deliberately have tried to erase (in the case of email) exists without their control.

Deploying this survey has been extremely useful in showing a) how little people know and understand about Google's personal user collection and b) that people *are* interested in a Google anonymizer. In response to the 85% of respondents who said they would be interested in a Google anonymizer, we have developed and present *Google-Anon.*

5.0 Google-Anon

Google-Anon is a web-based service we have created and made available for public consumption. This service is a tool that allows users to search Google anonymously and to see the differences between using Google Search anonymously and non-anonymously.

We feel that it is important that people are cognizant of what they are doing on the web and what aspects of their privacy are or are not as private as they may have thought. Our feelings are bolstered by the data garnered from our self-administered survey. Our survey clearly indicated that when most users use Google search, they are not aware of the information that Google tracks about them.

5.1 Project Goals

Our project was created in support of our thesis and with the following goals:

- Help people become more aware that Google collects and tracks data about them
- Demonstrate in real-time, the effects of Google specifying its services using the data it collects on users, specific to Google Search
- Show people the actual difference between an anonymous Google search experience and a non-anonymous Google search experience
- Provide an avenue for users to change their Google-service related habits by offering them a way to use Google's search engine anonymously

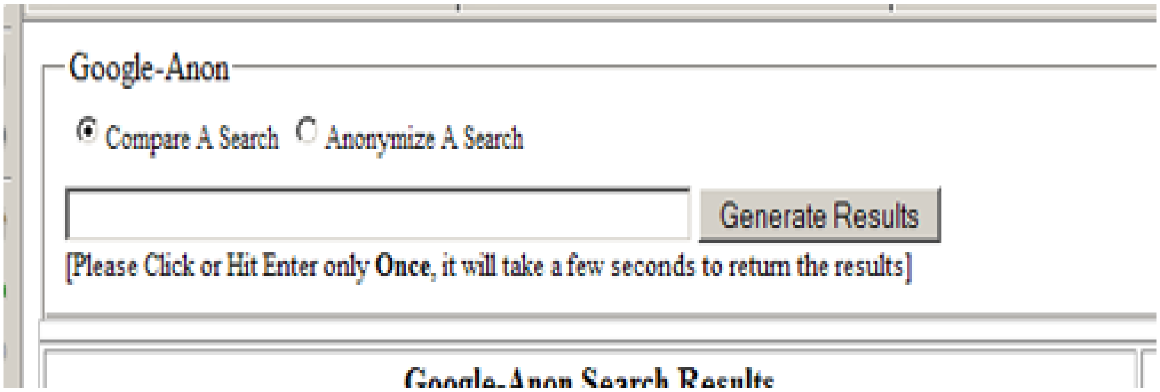
5.2 Project Description

"Google-Anon" is a fairly complex project that is based on the Python programming language, Javascript and HTML. Using some basic web-related development tools and a little logic we constructed a system that allows users to conduct a search, similar to the way they do in Google. Upon the input of a query by a user, our system goes through a set of steps in order to return results that are anonymously retrieved from Google search, with respect to the user, and in some cases these results are returned alongside results that are non-anonymously retrieved from Google search. In the case of our project, when anonymous results are displayed, certain areas of the returned results are emphasized to alert the user of information that would have been specifically targeted to their personal data if the search had not been anonymous.

5.2.1 Providing a User Interface

The first major step in our project was providing users with an acceptable user interface (UI). We chose to make a UI similar to the one that Google search displays in order to keep users familiar with how they should use our service for searching. At the same time we needed to allow the user to choose whether they wanted to perform a comparison search where they saw anonymous and non-anonymous results, or simply a lone anonymous search. Thus we provided the following components in a box at the top of our interface:

- 2 radio check boxes that allow the user to choose a comparison or solely anonymous search
- A text-input field that allows a user to enter search term(s) they would like to see results of
- A button that a user can click in order to initiate the search



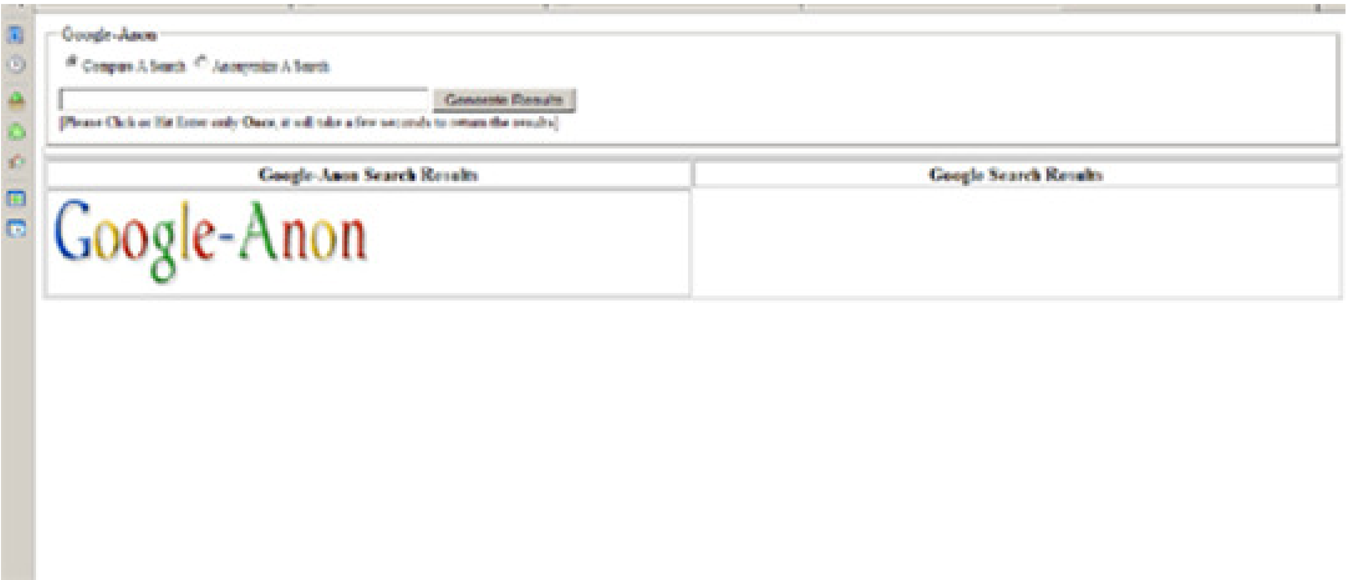
Google-Anon

● Compare A Search ○ Anonymize A Search

[Please Click or Hit Enter only **Once**, it will take a few seconds to return the results]

Figure 5.1 The user decides to "Compare a Search" or "Anonymize a Search", enters his/her query, and then clicks on "Generate Results" to initiate Google-Anon.

Below these components are simply headers and blank space that will contain the final results of the search when they are returned by the system.



Google-Anon

● Compare A Search ○ Anonymize A Search

[Please Click or Hit Enter only Once, it will take a few seconds to return the results]

Figure 5.2 Google-Anon's interface before user makes a query.

5.2.2 Processing a Google-Anon Search

After choosing a search type and entering search terms, a user clicks the "Generate Results" button to initiate a search. At this point Google-Anon must process this request to return the correct results. The following steps and details entail the processing of this request by our system:

First, the system grabs the input from the text-field and the radio-buttons. This data is accessed using javascript commands that can literally call for the data in those elements of the page by the following code command:

- `getElementById('nameOfElementDesired');`

Upon getting the data in the necessary fields, the system sends it to the server using an asynchronous javascript method. This method is commonly referred to as AJAX and allows the user to send data to the server without forcing the search page to reload.

When the server receives the data it creates the following string, which is a link that can be used to access Google's servers and make a search query:

- `http://google.com/search?q=<INPUT TEXT HERE>`

When this string is constructed, the server can literally open the url using the following method call, and request a page of Google search results:

- `urlopen('http://google.com/search?q=<INPUT TEXT HERE>');`

Before this method can be called though, the system needs to ensure that the request will be made anonymously. In order to achieve this two things are done:

1) an HTTPRequest header, a standard part of most requests made over the HTTP protocol on the internet, is created and attached to the request that will go out to Google servers.
2) A proxy server is contacted by the system and asked to request the url string we have created. This proxy server displays its own random IP address to Google servers, whilst asking for the search query the system passed to it, along with the header information that was created. In essence the proxy simply forwards our HTTPRequest, header, query-string and all.

The search initiated by the user has now been anonymously processed by the system and the results are being awaited.

5.2.3 Returning Anonymous and Regular Query Results

At this point in the system, if everything went as planned, the proxy has anonymously forwarded our request to Google and Google has returned the results of our request to the proxy, who has in turn returned the results to us. The system now has a string of text that represents the page that Google returned, which includes the results of our anonymous search. Further processing is now needed in order to return these results to the user as planned.

5.2.3.1 Anonymous Result Returning

In each case, the anonymous set of results is returned from the server back to the client's browser. Before doing this though, the system formats the results, so that two particular parts of the result page are emphasized. First the "# of results" section is highlighted with a green color, then the "Sponsored Links" section of the page is also highlighted with a green color. Both of these highlights are performed by setting properties in the html of the page using simple text-parsing techniques.

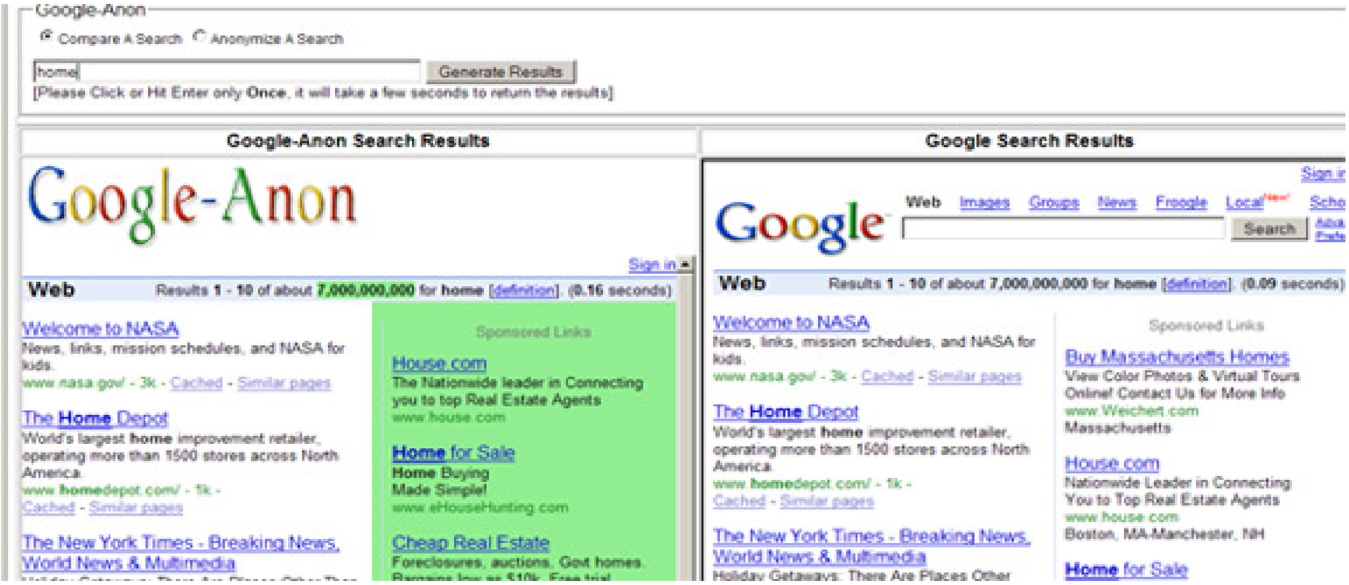
Once the page has been formatted correctly the system returns the page in string form, back to the user, where it will be handled by javascript that runs in the user's browser. Once the javascript receives the page, it displays it. Depending on whether the user chose a comparison search or a lone anonymous search, the javascript will display the anonymized results on the first half of the page or the entire page respectively.

5.2.3.2 Regular, Non-Anonymous Result Returning

The non-anonymous result of the user's search is processed and returned only if the user chooses to "Compare a Search" at the top of the page. Furthermore processing only occurs once the anonymous search has come back to the javascript in the user's browser, as mentioned above.

Upon the anonymous page results return, the javascript forces the user's browser to directly connect to Google's servers and request the search terms entered by the user. This is done by making the user's browser navigate to the same URL string that was constructed for the anonymous search ('http://google.com/searchq?=<INPUT TEXT HERE>'). The results of this connect and request are displayed in an IFRAME, also called inner frame, on the lone page of our service. The IFRAME is a component that acts as a browser within the page that our system gives to the user. The IFRAME allows its contents to connect to and display any URL that a regular browser would be able to display. Our IFRAME lies directly to the right of the anonymous results page that has been returned and takes up the other half of the page.

Below are some images to help clarify what is going on:



[Please Click or Hit Enter only **Once**, it will take a few seconds to return the results]

Results **1** - **10** of about **7,000,000,000** for **home** [definition]. (**0.16** seconds)

Welcome to NASA
News, links, mission schedules, and NASA for kids.
www.nasa.gov/ - 3k - Cached - Similar pages

The **Home** Depot
World's largest **home** improvement retailer, operating more than 1500 stores across North America.
www.**home**depot.com/ - 1k - Cached - Similar pages

The New York Times - Breaking News, World News & Multimedia

House.com
The Nationwide leader in Connecting you to top Real Estate Agents
www.house.com

Home for Sale
Home Buying Made Simple!
www.eHouseHunting.com

Cheap Real Estate
Foreclosures, auctions, Govt homes. Bargains low as $10k. Free trial

Results **1** - **10** of about **7,000,000,000** for **home** [definition]. (**0.09** seconds)

Welcome to NASA
News, links, mission schedules, and NASA for kids.
www.nasa.gov/ - 3k - Cached - Similar pages

The **Home** Depot
World's largest **home** improvement retailer, operating more than 1500 stores across North America.
www.**home**depot.com/ - 1k - Cached - Similar pages

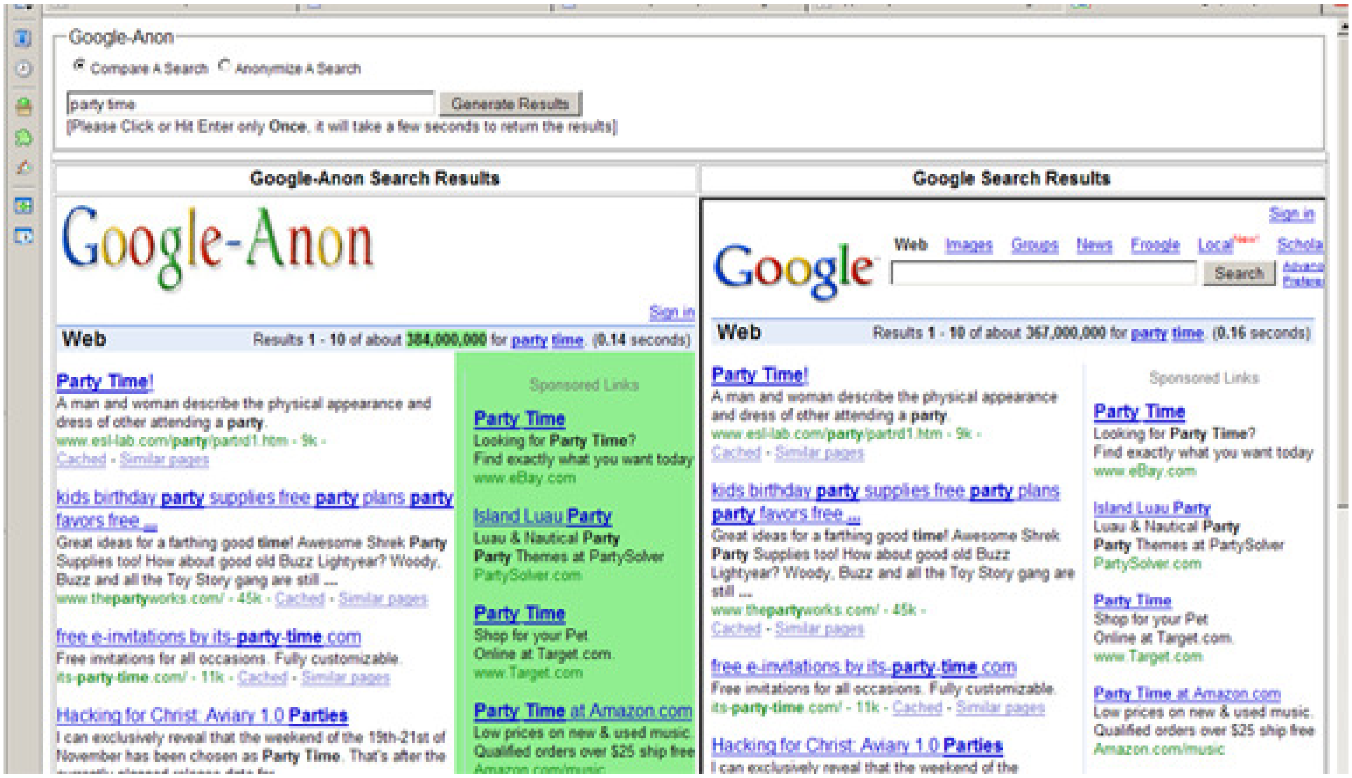
The New York Times - Breaking News, World News & Multimedia
Holiday Getaways: There Are Places Other

Buy Massachusetts Homes
View Color Photos & Virtual Tours Online! Contact Us for More Info
www.Weichert.com
Massachusetts

House.com
Nationwide Leader in Connecting You to Top Real Estate Agents
www.house.com
Boston, MA-Manchester, NH

Home for Sale

Figure 5.3 Results Page when user does a "Compare a Search" query of "home". Note the different ads on the side of each set of results. The ads are geographic-specific to the user.



Google-Anon

● Compare A Search ○ Anonymize A Search

party time

[Please Click or Hit Enter only **Once**, it will take a few seconds to return the results]

Google-Anon Search Results

Web Results **1** - **10** of about **384,000,000** for **party time**. (**0.14** seconds)

Party Time!
A man and woman describe the physical appearance and dress of other attending a **party**.
www.esl-lab.com/**party**/partrd1.htm - 9k - Cached - Similar pages

kids birthday **party** supplies free **party** plans **party** favors free ...
Great ideas for a farthing good **time**! Awesome Shrek **Party** Supplies too! How about good old Buzz Lightyear? Woody, Buzz and all the Toy Story gang are still ...
www.the**party**works.com/ - 45k - Cached - Similar pages

free e-invitations by its-**party-time**.com
Free invitations for all occasions. Fully customizable.
its-**party-time**.com/ - 11k - Cached - Similar pages

Hacking for Christ: Aviary 1.0 **Parties**
I can exclusively reveal that the weekend of the 19th-21st of November has been chosen as **Party Time**. That's after the

Party Time
Looking for **Party Time**?
Find exactly what you want today
www.eBay.com

Island Luau **Party**
Luau & Nautical **Party**
Party Themes at PartySolver
PartySolver.com

Party Time
Shop for your Pet
Online at Target.com.
www.Target.com

Party Time at Amazon.com
Low prices on new & used music.
Qualified orders over $25 ship free
Amazon.com/music

Google Search Results

Web Results **1** - **10** of about **367,000,000** for **party time**. (**0.16** seconds)

Party Time!
A man and woman describe the physical appearance and dress of other attending a **party**.
www.esl-lab.com/**party**/partrd1.htm - 9k - Cached - Similar pages

kids birthday **party** supplies free **party** plans **party** favors free ...
Great ideas for a farthing good **time**! Awesome Shrek **Party** Supplies too! How about good old Buzz Lightyear? Woody, Buzz and all the Toy Story gang are still ...
www.the**party**works.com/ - 45k - Cached - Similar pages

free e-invitations by its-**party-time**.com
Free invitations for all occasions. Fully customizable.
its-**party-time**.com/ - 11k - Cached - Similar pages

Hacking for Christ: Aviary 1.0 **Parties**
I can exclusively reveal that the weekend of the

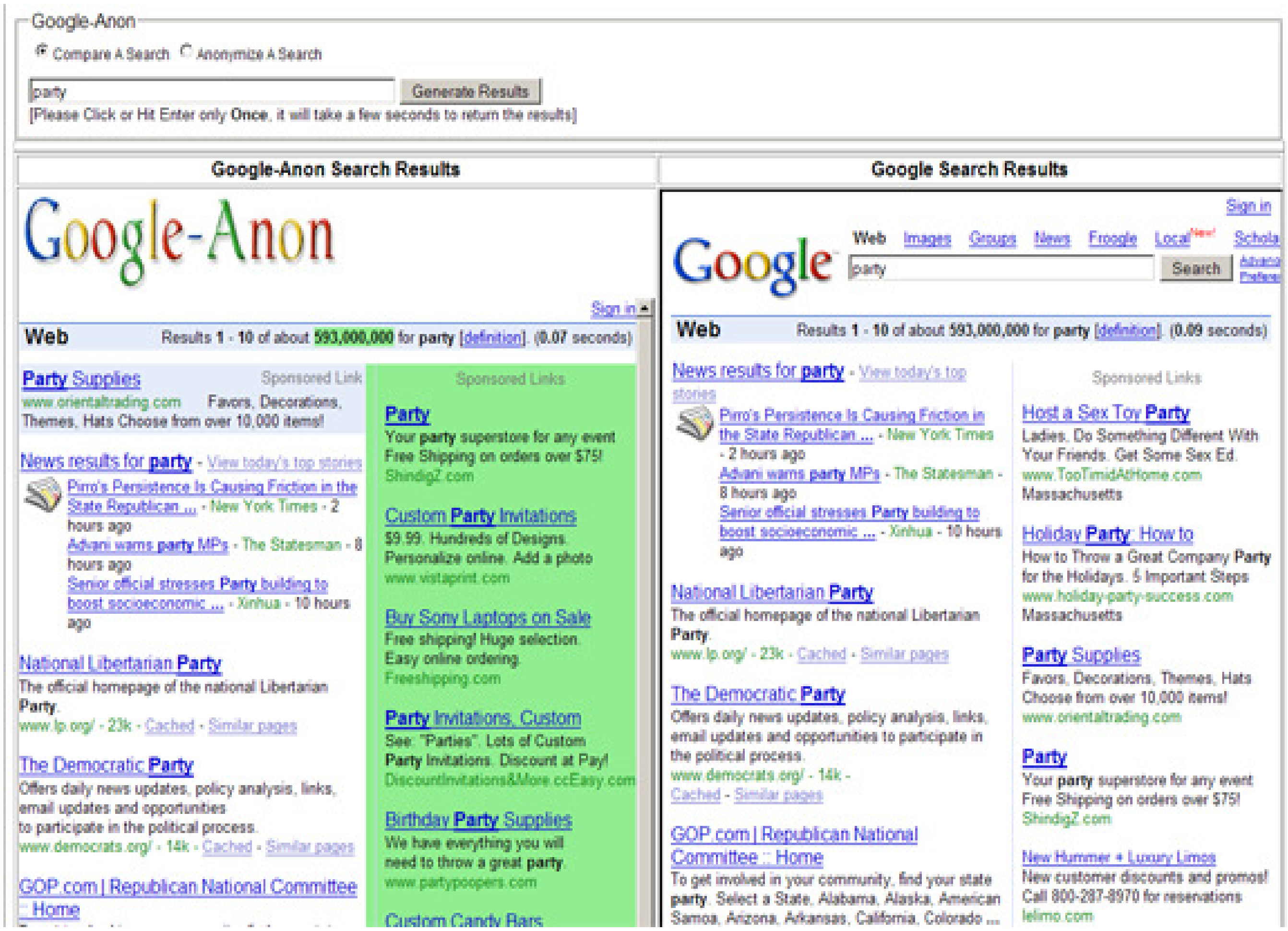
Party Time
Looking for **Party Time**?
Find exactly what you want today
www.eBay.com

Island Luau **Party**
Luau & Nautical **Party**
Party Themes at PartySolver
PartySolver.com

Party Time
Shop for your Pet
Online at Target.com.
www.Target.com

Party Time at Amazon.com
Low prices on new & used music.
Qualified orders over $25 ship free
Amazon.com/music

Figure 5.4 Results page of a "Compare a Search" query of "party time". Notice the difference in number of results between the anonymized and unanonymized searches


Google-Anon
○ Compare A Search ○ Anonymize A Search
party
Generate Results

[Please Click or Hit Enter only Once, it will take a few seconds to return the results]

Google-Anon Search Results
Google-Anon
Sign in
Web

Results 1 - 10 of about 593,000,000 for party [definition]. (0.07 seconds)

Party Supplies — Sponsored Link
www.orientaltrading.com Favors, Decorations, Themes, Hats Choose from over 10,000 items!

News results for party - View today's top stories
Pirro's Persistence Is Causing Friction in the State Republican ... - New York Times - 2 hours ago
Advani warns party MPs - The Statesman - 8 hours ago
Senior official stresses Party building to boost socioeconomic ... - Xinhua - 10 hours ago

National Libertarian Party
The official homepage of the national Libertarian Party.
www.lp.org/ - 23k - Cached - Similar pages

The Democratic Party
Offers daily news updates, policy analysis, links, email updates and opportunities to participate in the political process.
www.democrats.org/ - 14k - Cached - Similar pages

GOP.com | Republican National Committee :: Home

Sponsored Links

Party
Your party superstore for any event Free Shipping on orders over $75!
ShindigZ.com

Custom Party Invitations
$9.99. Hundreds of Designs. Personalize online. Add a photo
www.vistaprint.com

Buy Sony Laptops on Sale
Free shipping! Huge selection. Easy online ordering.
Freeshipping.com

Party Invitations, Custom
See: "Parties". Lots of Custom Party Invitations. Discount at Pay!
DiscountInvitations&More.ccEasy.com

Birthday Party Supplies
We have everything you will need to throw a great party.
www.partypoopers.com

Custom Candy Bars

Results 1 - 10 of about 593,000,000 for party [definition]. (0.09 seconds)

News results for party - View today's top stories
Pirro's Persistence Is Causing Friction in the State Republican ... - New York Times - 2 hours ago
Advani warns party MPs - The Statesman - 8 hours ago
Senior official stresses Party building to boost socioeconomic ... - Xinhua - 10 hours ago

National Libertarian Party
The official homepage of the national Libertarian Party.
www.lp.org/ - 23k - Cached - Similar pages

The Democratic Party
Offers daily news updates, policy analysis, links, email updates and opportunities to participate in the political process.
www.democrats.org/ - 14k - Cached - Similar pages

GOP.com | Republican National Committee :: Home
To get involved in your community, find your state party. Select a State, Alabama, Alaska, American Samoa, Arizona, Arkansas, California, Colorado ...

Sponsored Links

Host a Sex Toy Party
Ladies. Do Something Different With Your Friends. Get Some Sex Ed.
www.TooTimidAtHome.com
Massachusetts

Holiday Party: How to
How to Throw a Great Company Party for the Holidays. 5 Important Steps
www.holiday-party-success.com
Massachusetts

Party Supplies
Favors, Decorations, Themes, Hats Choose from over 10,000 items!
www.orientaltrading.com

Party
Your party superstore for any event Free Shipping on orders over $75!
ShindigZ.com

New Hummer + Luxury Limos
New customer discounts and promos! Call 800-287-8970 for reservations
lelimo.com

Figure 5.5 Results Page when user does a "Compare a Search" query of "home". Again, note the different ads on the side of each set of results.

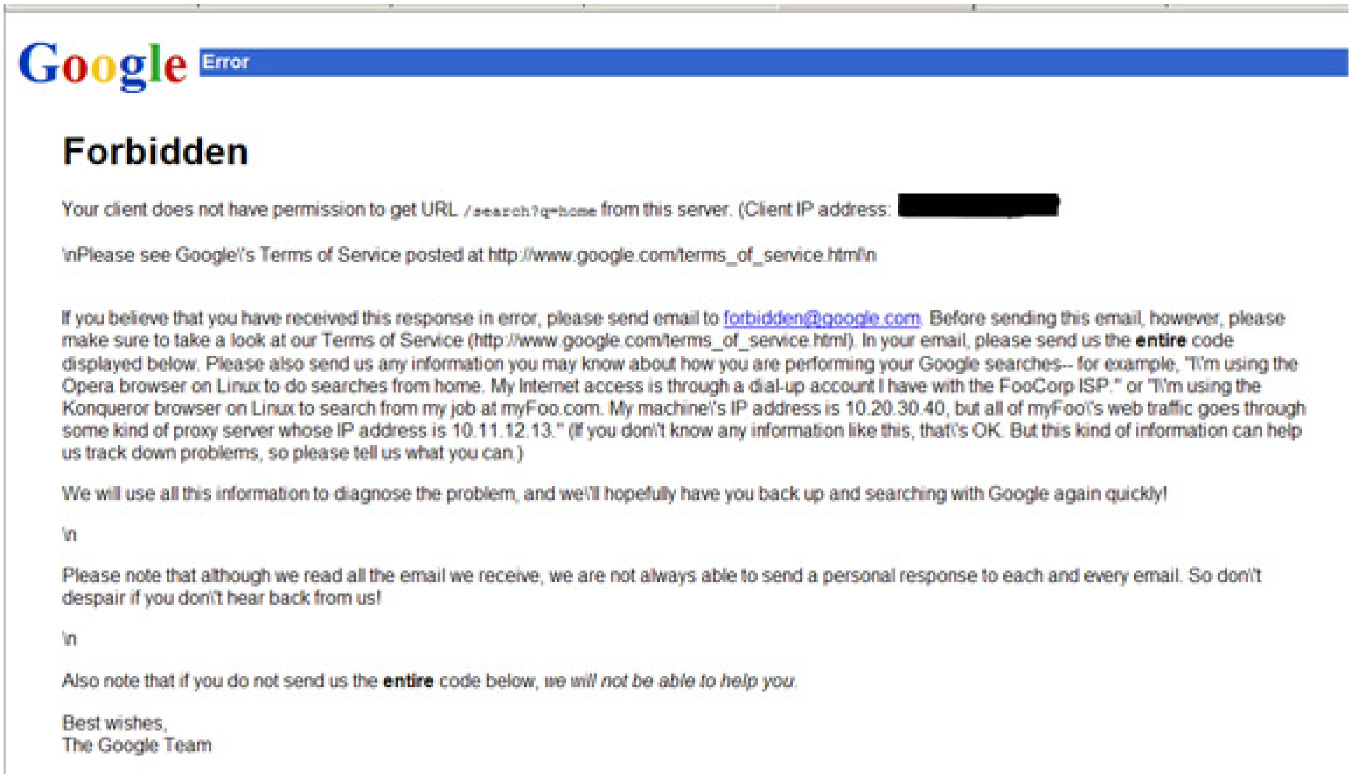
5.3 Project Challenges and Issues

In creating our system, we faced a number of challenging obstacles. On the one hand, we wanted to create a service that allows users to make anonymous searches as fast or comparably as fast as Google searches. On the other hand, Google is much more clever than expected by virtue of its brilliant employees, and that being said there are a number of things Google does to "prevent" people from performing certain "adverse" activities related to their services.

5.3.1 Making Users Anonymous

The heart of our project is making users anonymous. In order to do this we decided that we, not the user would have to make the request from Google for search results that a user wanted.

Our first attempt at acting as a "proxy" resulted in a flat denial right off the bat. When you try to simply write a line of code to open up the URL string that gives Google search results, you get a response message like the one below that tells you that you are not able to make the request.



Forbidden

Your client does not have permission to get URL /search?q=home from this server. (Client IP address: [redacted]

\nPlease see Google\'s Terms of Service posted at http://www.google.com/terms_of_service.html\n

If you believe that you have received this response in error, please send email to forbidden@google.com. Before sending this email, however, please make sure to take a look at our Terms of Service (http://www.google.com/terms_of_service.html). In your email, please send us the **entire** code displayed below. Please also send us any information you may know about how you are performing your Google searches-- for example, "I\'m using the Opera browser on Linux to do searches from home. My Internet access is through a dial-up account I have with the FooCorp ISP." or "I\'m using the Konqueror browser on Linux to search from my job at myFoo.com. My machine\'s IP address is 10.20.30.40, but all of myFoo\'s web traffic goes through some kind of proxy server whose IP address is 10.11.12.13." (If you don\'t know any information like this, that\'s OK. But this kind of information can help us track down problems, so please tell us what you can.)

We will use all this information to diagnose the problem, and we\'ll hopefully have you back up and searching with Google again quickly!

\n

Please note that although we read all the email we receive, we are not always able to send a personal response to each and every email. So don\'t despair if you don\'t hear back from us!

\n

Also note that if you do not send us the **entire** code below, *we will not be able to help you.*

Best wishes,
The Google Team

Figure 5.6 When attempting to retrieve Google's Search Results programmatically, Google returns the shown page.

After feeling outwitted by Google, we finally realized that our system would have to imitate a browser in order to get Google to acknowledge our search request. In order to do this from within the code we had to set an HTTPRequest header variable when we made our request to Google (as described above). This request header was sent with our Google query and thus allowed us to get some results back until we hit another roadblock.

5.3.2 Getting Past Google's Cleverness

Our next challenge would prove to be the largest as we attempted to get past more Google cleverness and our own inexperience with web-based services.

It started when we realized that whilst the user may be anonymous, we as a server were not anonymous and would not return interesting or different results for users as they used our service. Furthermore Google might feel that it no longer wanted to service requests from our system IP address and block us from using the service. This would render us completely helpless. At this point we decided to use some knowledge given to us by MIT's Network Admin Jeff Schiller and use a system called the Tor network. This system was a high-anonymity network that would allow people to make requests from certain pages and return the answers to those requests undetected. The Tor network is short for “The onion routing” network and is called so because it uses a multi-hop system for routing and returning network requests. When a user makes a request through the Tor system, that request is forwarded through three different machines in the Tor network and the response/answer is returned along that same route.

We installed the Tor software and began using it as an access point for our searches. All seemed well until we began getting some use on the system. We found out two major issues that would lead to us abandoning the benefits of the Tor network:

1) Requests were too slow. Because the request for Google search results had to travel between three different machines that were possibly spread across the globe, and then perform that same trip on returning a response, our response time to users was entirely too slow to make it an alternative to using Google search.
2) The Tor software used the same three-hop route for at least 10 minutes when we made a request. This actually resulted in Google blocking our end-route-IP address when we had light usage on the server because it interpreted all the requests from one IP address as an attack on its system. This made our system unusable to more than one person at a time which was unacceptable.

Unable to use the Tor network, we took a page from that same playbook and decided that we should and would have to use some sort of proxy system when making requests from our server. We also reasoned that we could not simply use one proxy for a period of time since that IP would eventually be blocked as it was when we were using Tor network. Our solution was to create a system of randomly rotating proxy servers that would be used by our server. The plan was that every time a user made a request using Google-Anon we would randomly retrieve a proxy from a large list of available proxies, and ask the retrieved proxy to make the request on our behalf. Ironically the hardest part of this task was finding a list of proxy servers that were, public, reliable, and allowed direct connection without physical human presence. The effort of finding a good listing of publicly available, programmatically accessible proxy servers is underrated. Since the new system of randomly rotated proxies has been in place, we have not had any major problems with our service.

The occasional hiccup may occur in the system, but is quickly fixed by a browser reload. We attribute these minor issues to a beta prototype, as well as a lack of enterprise level software for our server.

We have put Google-Anon on the MIT network and anticipate that its use will spread via word of mouth. We will collect and process data about the site (e.g. how many visitors use Google-Anon, how many use the anonymizer alone, how many use the comparison search, etc) only to find out if more people are interested in such services. From there, perhaps we will create or inspire others to create more anonymizing techniques for other Google services.

6.0 Conclusions

This paper presents a mere glimpse of what kind of information Google collects about its users and what can happen to that information. We discussed Google's most pervasive services, Google Search, AdSense, and Gmail and the type of user profiles Google can create with the information collected from these services. However, we ignored many of Google's smaller services, such as Desktop, Toolbar, and Analytics that are gaining popularity and further intrude on users' privacy. Factoring in these services, the amount of information presented in the "Bob Smith" profile of Section 2.4 would be augmented to include even more penetrating information, such as Bob Smith's desktop activity and comprehensive internet activity . As these intrusive services grow in popularity, they will ensure Google's ability to gain an even stronger hold on people's personal information.

Despite Google's growing hold on users' data, the survey in Section 4 shows that most Google users remain oblivious to the amount of personal user information Google collects. The survey also shows that when presented with an anonymizing alternative, people will take interest in such an option. People need to recognize what information is being gathered about them, and with that knowledge decide if and to what extent they want to protect their privacy.

While Google's data management seems daunting, Google is not the first and certainly will not be the last company to acquire and store user data on a major scale. As society hurtles toward an era where user data collection is the norm, we must answer an inevitable question about privacy: Should we continue to fight for our personal privacy or should we accept this transparency as a natural progression of technology?

In order for society to make this decision, people must first recognize the issue and then be presented with the appropriate tools to deal with it. We hope with this paper, the presented survey, and Google-Anon, will initiate and bring light to a necessary discussion about Google, its privacy issues, and the general progression of user data transparency.

7.0 References

[1] http://www.google.com/corporate/history.html (visited December 5, 2005).

[2] http://www.google.com.au/press/timeline.html (visited December 5, 2005).

[3] http://desktop.google.com/about.html (visited December 5, 2005).

[4] http://wwww.google.com/support/toolbar/bin/answer.py?answer=14292&topic=938 (visited December 5, 2005).

[5] http://internetstockblog.com/article/4938 (visited December 13, 2005).

[6] http://www.google.com/search?hl=en&q=the&btnG=Google+Search, *Google search of "the" produced 9.1 billion results,* (visited December 5, 2005).

[7] The Anatomy of a Large-Scale Hypertexual Web Search Engine;" Brin, Sergey; Page, Lawrence; http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm; (visited December 5, 2005).

[8] http://www.google.com/technology/index.html (visited December 5, 2005).

[9] http://www.google.com/privacy_faq.html (visited December 5, 2005).

[10] http://www.thisearly.com/content/view/93/2 (visited December 5, 2005).

[11] http://en.wikipedia.org/wiki/IP_address (visited December 5, 2005).

[12] http://www.iana.org/ipaddress/ip-addresses.htm (visited December 5, 2005).

[13] http://en.wikipedia.org/wiki/Internet_Assigned_Numbers_Authority (visited December 5, 2005).

[14] http://computer.howstuffworks.com/nat.htm/printable (visited December 12, 2005).

[15] http://whatismyip.com (visited December 12, 2005).

[16] http://www.whois.sc/ (visited December 12, 2005).

[17] http://www.google.com/services/adsense_tour/page2.html (visited December 5, 2005).

[18] https://www.google.com/adsense/new (visited December 5, 2005).

[19] https://adwords.google.com/support/bin/answer.py?answer=6382&hl=en (visited December 5, 2005).

[20] http://gmail.google.com (visited December 5, 2005).

[21] https://www.google.com/accounts/SmsMailSignup1 (visited December 5, 2005).

[22] http://mail.google.com/mail/help/privacy.html (visited December 5, 2005).

[23] http://www.google.com/privacypolicy.html (visited December 5, 2005).

[24] http://www.export.gov/safeharbor/ (visited December 5, 2005).

[25] http://www.export.gov/safeharbor/sh_overview.html (visited December 5, 2005).

[26] http://en.wikipedia.org/wiki/Patriot_act (visited December 5, 2005).

[27] http://www.wral.com/news/5287261/detail.html (visited December 5, 2005).

[28] http://www.google-watch.org/krane.html (visited December 5, 2005).