Googling Considered Harmful

Gregory Conti
United States Military Academy
West Point, New York
conti@acm.org

ABSTRACT

Virtually every Internet user on the planet uses the powerful free tools offered by a handful of information service providers in many aspects of their personal and professional lives. As a result, users and organizations are freely providing unprecedented amounts of sensitive information in return for such services as Internet search, email, mapping, blog hosting, instant messaging and language translation. Traditional security measures, such as cryptography and network firewalls, are largely ineffective because of the implicit trust paradigm with the service provider. In this paper, we directly address this problem by providing a threat analysis framework of information disclosure vectors, including fingerprinting of individuals and groups based on their online activities, examine the effectiveness of existing privacy countermeasures and clearly outline the critical future work required to protect our corporate, organizational and individual privacy when using these services.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval, K.4.2 [**Computers and Society**]: Social Issues, K.4.3 [**Computers and Society**]: Organizational Impacts, K.4.4 [**Computers and Society**]: Electronic Commerce.

General Terms

Security, Legal Aspects, Human Factors

Keywords

googling, Google, privacy, anonymization, usable security, information disclosure, anonymity, fingerprinting, search, AOL

1. INTRODUCTION

The innocuous and ubiquitous web search box presents a significant and increasing risk to individuals and organizations. The information that we disclose over time provides to the information service provider a progressively clearer picture of our personal and professional lives, the lives of our associates and the health, stratagems and structure of our private and public organizations. While the near frictionless flow of information is key to the success of today's Internet, it is also an ever-increasing threat. The low cost of high capacity storage devices combined with business models that depend on customized advertising and content virtually ensure that the information we provide will never be discarded. During the initial Darwinian years of the World Wide Web, the diversity and short lifespan of web-based information service companies diffused our personal information, but the recent rise of a handful of companies to industry dominance places extremely large portions of this information, and hence exceptional power, into a relatively few hands. The problem goes beyond simple web search; each compelling, and often free, new service and tool offered by these information providers exacerbate the problem by increasing the flow of information. Counterintuitively, traditional improvements such as increased network bandwidth and enhanced usability worsen the problem by attracting more users and facilitating information leakage. Globally, Internet users (and frequently their employers) generally accept the paradigm that such usage is safe,[*] that privacy policies provide adequate protection and that nearly unqualified trust in information service providers is merited. They are essentially unaware, or unconcerned, that, over time, the sum total of their interactions paints a portrait of their political, economic and social situations of unprecedented proportions. The August 2006 accidental disclosure by America Online (AOL) of the search activity of 650,000 AOL users dramatically underscores the problem.

With approximately one billion Internet users worldwide [1] and the potential for permanent storage of our interactions, dialog on this problem needs to move forward and potential solutions developed. Based on these concerns, the purpose of this paper is to better define and motivate the problem, discuss why current personal and corporate privacy countermeasures are inadequate, present potential solutions and lay the groundwork for necessary future work. Our work is based exclusively on publicly available sources and not on any insider or otherwise privileged information.

The results we present in this paper can be generalized to any information production and consumption relationship, but we focus our analysis on one of the current industry leaders and

New Security Paradigms Workshop '06, September 18-21, 2006, Schloss Dagstuhl, Germany.

[*] Alternatively, more security minded users will likely consider the risk/convenience trade-off of using a given service. Many will ultimately decide to trust the information service provider with some amount of sensitive information in return for the convenience of the service. As with other types of security, if it gets in the way of doing your job, you're likely to opt for the convenience.

arguably the world's largest search engine [2], Google. While we assume that Google is a non-malicious entity, they do face a legal requirement to act in the best interests of their shareholders. This fact creates a tension between making a profit and their stated goal of providing long-term value for their end users [3] as well as their informal corporate motto "Don't be evil." [4] It is important to note, despite our assumption, that there have been incidents that call into question their "Don't be evil" philosophy. The following are two examples. The first concerns censorship. Initially, Google's policies stated that "Google does not censor results for any search term," [5] but in early 2006 they changed their policy and began censoring some content at the behest of the Chinese government [6]. The second is apparent retribution for unfavorable media coverage. In July 2005, the CNET media outlet published an article by Elinor Mills containing sensitive personal information on the CEO of Google [7]. The information was collected using Google tools. Shortly thereafter, it was reported that Google would not talk to CNET reporters for one year [8]. While Google, by choosing "Don't be evil" as their corporate motto has admirably set a very high standard for their company, the fact remains that evil is subjective. In the words of Google CEO Eric Schmidt "Evil is what Sergey[†] says is evil."

Google proclaims its mission "is to organize the world's information and make it universally accessible and useful." [2] By all accounts they take this mission very seriously. While exact usage statistics are not publicly available, Google states that their user base "is in the millions" [9] and a recent media report estimates that they receive 380 million visitors each month [10]. As of 2003, they reportedly responded to 250 million search queries per day [11] and as of January 2006 they have indexed 9.7 billion web pages, 1.3 billion images and more than one billion Usenet messages [12]. Google has 138 country specific web portals in approximately 116 languages [13].[‡] They offer 21 free web services and 11 free tools [14]. [§] There are many more brewing in Google Labs, which currently lists 16 publicly disclosed projects [15]. The number of additional projects being developed using Google's highly publicized "20 percent time" with which Google engineers are "free to pursue projects they are passionate about" [16] is not publicly known.

Information does not just flow from individual users, companies and governments to Google, but also from third party[**] web activity as well as information harvested directly by Google's Googlebot web spider [17]. In many instances, information can be clustered at the organizational level due to publicly available Internet Protocol (IP) address allocations [18] and geolocation data [19]. If, given our assumption, this information is unlikely to be discarded, the sum will therefore increase over years and decades, perhaps beyond the life of the individual user or organization. Current technology and Google expertise in data mining, artificial intelligence, genetic algorithms, information retrieval, machine learning, natural language processing and profiling provide the capability to deeply explore and interconnect the data. (Note that experts in all of these research domains are currently being sought by Google Labs [15].) Insights that are not possible using today's technology are likely to be feasible over the next decades. Even today Google is thought capable of keeping the entire Internet in RAM. In the future they may be capable of archiving continuous historical snapshots of its state [20,21]. When aggregated individual and organizational data is combined with Google's top tier intellectual talent and world-class information processing resources it arguably gives them the information resources of a nation-state and constitutes a significant threat if not properly managed. Consider the following scenarios that illustrate the potential risks.

- A Gmail user experiences a death of a parent. She uses her Gmail account to inform all her family and friends. Google is now aware of her social network and also adds a bereavement counseling advertisement to the email.

- A company is facing undisclosed financial difficulties known only to company insiders. Employees are concerned. Google searches may show a surge in job search activity from the company's IP address range.

- An anonymous individual runs a blog, discussion board and support group for people afflicted with a serious medical condition. Other Google web users (third parties) frequently search for the blog and use the individual's full name in the search, tying the two together.

- A company mandates the use of the Google Desktop application. At some point in the future, a vulnerability is discovered by computer criminals. The entire company network is now at risk.

- A security expert gives a talk on utilizing Google to search for vulnerable computers on the Internet. Google responds in a way that prevents end users from conducting such searches, but Google itself still has the capability to conduct the searches using internal resources.

- Anywhere on the planet, every time a user, with their homepage set to Google, opens their browser, Google is aware of the IP address.

- Google Maps satellite data is upgraded to higher resolution and receives broad attention due to being posted on a popular technology news blog. Users across the globe use the tool to examine their home, the homes of their friends and family as well as their place of employment.

[†] Sergey Brin is the cofounder of Google.

[‡] At least three of these languages are intended to be humorous.

[§] These counts include only major services, tools and tool packages. The actual number of all applications is far greater.

[**] Third party web activity includes an outside individual providing information about a person or organization. For example, a hacker may carefully separate an online pseudonym from their real world identity. Despite this care, an acquaintance might link the two identities together by searching for both the pseudonym and real world name in an attempt to locate the hacker's webpage.

- A company is considering the purchase of a new security system. Google searches from their network IP address block indicate the specific products under consideration.

As we look to the future and consider possible scenarios, we must consider probable advances in science. Given the resources that Google commands, it is difficult to determine what advances will result. Fingerprinting of network users is one likely outcome. Google's business model is based upon targeted advertising and it is probable that they will seek to tie clusters of interactions, across multiple computing platforms, with specific individuals and organizations. This situation is analogous to the use of encryption. Cryptographers typically consider encryption valid for only a period of time due to cryptanalytic advances and increased processing power. Likewise, we should assume our anonymity is only a function of time. Eventually, given enough information disclosing interactions, privacy will be compromised. Current anonymization countermeasures increase the time required for fingerprinting, but ultimately will only delay the inevitable when faced with Google's capabilities.

The mere existence of an information stockpile of this magnitude and potential lifespan presents a significant risk that can no longer be ignored. Even with our assumption that Google is non-malicious, there are significant threats despite Google's best efforts to the contrary. As is the case with any individual or organization, Google is fallible. For example, Google provides a number of free desktop applications. If one of these applications were found to possess a security flaw, a large number of users would be at risk.[††] Unfortunately, this is already the case [22,23]. Google's privacy policies are not sufficient protection, because they are malleable. In Google's parlance they "may change from time to time." [24] Such was the case with our earlier Chinese censorship example. The recent broad ranging subpoena by the United States Department of Justice for Google search records [25] demonstrates another issue. Despite Google's best intentions of keeping our transactions private, they are still subject to laws which can compel them to provide certain information. There is also the possibility of eavesdropping at any point in the network path from the end user's computer to Google's servers as most search transactions are sent unencrypted. See Kaufman [26] for an excellent overview of possible eavesdropping attacks. While web based encryption techniques such as Secure Sockets Layer (SSL) provide some protection against eavesdropping, the fundamental problem of trust remains. SSL only provides protection between the user's computer and Google's servers. Google is a trusted party in the interaction and is hence able to decrypt the communication. Media reports indicate that Google performs rigorous background checks of its employees and contractors, but the potential for malicious insiders always exists [27]. Consider the Robert Hansen case. Hansen, a trusted Federal Bureau of Investigation agent, was convicted of spying for the Soviet Union [28]. Similarly, Aldrich Ames, a 31-year veteran of the Central Intelligence Agency with a top-secret security clearance, was sentenced to life in prison for spying on behalf of a foreign power [29]. These incidents demonstrate that background checks and rigorous internal controls are not a panacea. Accidental disclosure of private information is another very real threat. In the first three months of 2006 alone, there were at least 50 large-scale data disclosure incidents potentially affecting more than 21.3 million people in the United States [30].[‡‡]

In this paper, we make the following contributions. We provide a framework for analysis of information service providers and use it, with Google as the case study, to examine two key aspects of the problem: information disclosure and fingerprinting of network activities. The result is a comprehensive threat analysis of Google, from both the end user and organizational perspectives. After this analysis, we examine the extent that existing privacy countermeasures can help mitigate the threats we have identified. Finally, we clearly outline future work that is essential to combating the problem.

The uniqueness of this work springs from our comprehensive threat analysis of information services and tools. By using Google as an in depth case study we demonstrate that organizations and individuals, over time, are trading tremendous amounts of sensitive information in return for free tools and services. The rate at which information is disclosed falls below the user's detection threshold and results in a general lack of awareness of the magnitude of the total disclosure. The recent Chinese censorship and Department of Justice court cases have raised awareness to some degree, but primarily have resulted in limited discussion, mainly in trade magazines, blogs and personal websites, that highlight isolated components of the problem. See [31,32,33] for representative examples. Of these, the best, albeit high-level, analysis can be found in Mills' controversial CNET article [7]. Similarly, organizations like the Electronic Frontier Foundation (EFF) [34,35,36], the Electronic Privacy Information Center (EPIC) [37] and Search Engine Watch [38] have provided analysis, primarily on policy based aspects of the problem. In this paper, we provide an overarching analysis and technical detail that ties these components together. Anonymization countermeasures exist, most notably Tor [39], Anonymizer [40] and Crowds[41], but they lack widespread adoption. Aimed primarily at keeping data available for a finite period of time, techniques such as the Ephemerizer server [42] bear promise in providing information service providers robust tools for safely destroying information within their organizational specific policy constraints. Zero knowledge protocols [43] and oblivious transfer techniques [44] also show promise, but have not been widely adopted. Likely driven by business requirements, network based user-profiling [45,46] serves to decrease anonymity and exacerbates the overall problem. Finally, usable security researchers have also addressed individual aspects of information disclosure [47], most notably phishing attacks. These works address only information gathered by a malicious adversary, not, from the end user's perspective, a trusted party such as a web search engine.

[††] Clearly, this situation occurs with other software manufactures as well, most notably the popular Microsoft Windows operating system.

[‡‡] These statistics are based on public announcements and media reports. We believe the actual number is at least an order of magnitude greater as historically many organizations hesitate to disclose such events due to fear of tarnished reputations and loss of trust with clients and business partners.

The rest of the paper is organized as follows. In Section 2 we perform an information disclosure-based threat analysis of Google's and other organization's tools and services. In Section 3 we examine user and organizational fingerprinting that will tie together disclosed information with specific entities. In Section 4 we examine existing countermeasures and discuss their relative effectiveness at reducing the threat. In Sections 5 and 6, we present critical areas for future work and our conclusions.

2. The Information Disclosure Threat

Every interaction we have with an information service provider discloses some information. For purposes of our analysis, in this section, we will focus on the disclosure of content. We define content as information we *deliberately* release to the information service provider, such as emails, web search terms and chat. We will consider other information, often passively released by the user, such as browser cookies, network packet headers, and browser environment variables in Section 3.

As we begin our discussion, consider typical interactions with the services depicted in Figure 1. We've placed these services on the axis based roughly on the amount of information that they receive from users. Any similar service may be placed somewhere on the line. Actual placement will vary based upon individual and organizational usage habits. The graph will change over time. Given our assumption that information is never discarded; services will remain at constant positions if they are no longer used or will shift to the right at a rate proportional to continued information disclosure (and hence continued use). Figure 2 expands this notion by incorporating time. It depicts the information disclosed via four typical online services: email, web search, Internet telephony and instant messaging. Depending on individual or organizational usage habits, the slope of the curve will vary, but will never be negative, as information is never discarded. The only exception is when a service is no longer used and then the result is a horizontal line. If a single information service provider offers multiple services, the total information disclosed to that organization is the sum of all of its individual services, see Figure 3. Note that the total information disclosure may also be increased via sharing of information with another information service provider, possibly due to a cooperative agreement, acquisition or merger.

While the figures represent theoretical information disclosure, our actual disclosures are diverse, significant and sensitive, particularly when aggregated over time. During initial pilot research, we examined the contents of Firefox's form field caches. While the cache file format does not distinguish data disclosed by destination website, the cumulative amount of information stored in these files surprised us with its depth and breadth. We asked several users to identify the number of items found in their cache which they consider to be sensitive, e.g. they did not wish to publicly share the search queries. Our initial results ranged from 5% for a user on an employer monitored corporate network to 34% for a user's home web surfing. The problem is far worse when we consider not just search, but a more complete offering of information services. (See Table 1) We discuss tying the disclosure of this information to specific individuals and organizations in the next section, but leave additional analysis of real world user data for future work.

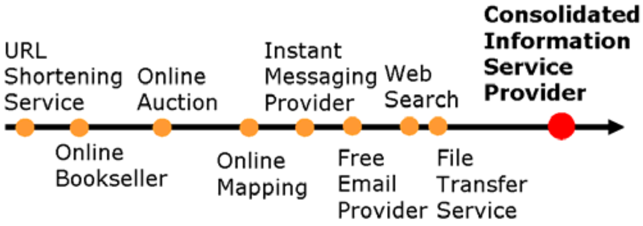

Figure 1: Information Disclosure Axis: A Comparison by Provider. Actual placement on the axis depends on a given individual's or organization's activities. Over time, positions will slide to the right if more information is disclosed. They will remain constant if the service is no longer used.

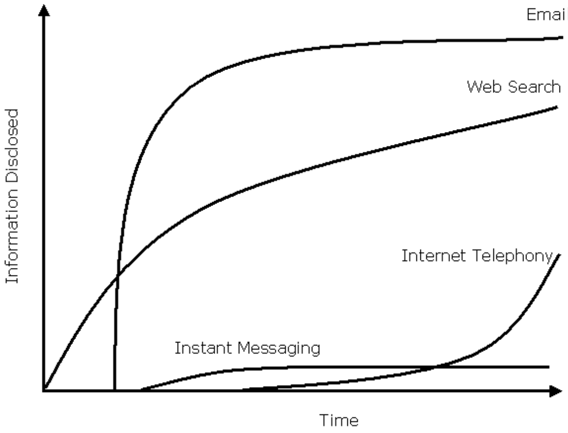

Figure 2: Information Disclosure Over Time: Unique to each individual's or organization's online activities, this figure depicts information disclosure as new services are adopted and older services (Instant Messaging in the figure) are discarded. Note that information disclosed never declines.

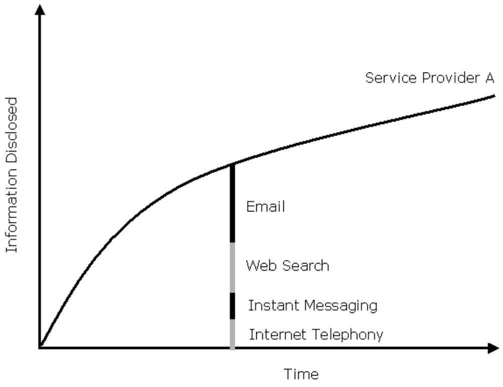

Figure 3: Instantaneous Snapshot of Information Disclosure by Service. By using services provided by a single organization, the total information disclosure is the sum of all service interactions at a given time.

Table 2: Partial Listing of Google Services and Possible Information Disclosure Threats

Service	**Description**	**Possible Information Disclosure**
Alerts	News alerts via email	news stories and topics you are interested in
Answers	Paid research service	research interests
Blog Search	Blog search engine	work-related and personal interests
Blogger	Blog hosting and search	work-related and personal interests
Book Search	Full text book search	work-related and personal interests
Calendar	Multi-user online calendaring service	time, date and location of your engagements
Catalogs	Mail order catalog search and browsing	items you may purchase
Code	API's and open source code	coding expertise
Compute	Donate idle computer time for research purposes	possible exposure of entire computing platform's stored data
Currency Conversions	Perform currency conversions	possible travel locations and business partners
Definitions	Glossary definitions	words you do not know the meaning of
Desktop	At a glance access to personal information	possible exposure of entire computing platform's stored data
Directory	Topic based web directory	work-related and personal interests
Earth	Location specific mapping, search and satellite imagery	possible exposure of entire computing platform's stored data
Finance	Business information and news	investment plans
Froogle	Online shopping	items you may purchase
Gmail	Web-based email service w/2GB storage	email addresses of your social network
Gmail for your Domain	Email hosting	email of all users of your domain
Groups	Mailing lists and discussion groups	work-related and personal interests
Images	Image search of web	work-related and personal interests
Local	Location based business and services search	current location, possible dining plans
Maps	Mapping and directions	location of friends and family, special events
Local for Mobile	Location specific searches	work-related and personal interests
Movies	Movie reviews and show times	movie interests, location, possible movie attendance
Music Search	Search wide range of music information	musical interests
News	News search and headlines	work-related and personal interests
Pack	Free collection of software	possible exposure of entire computing platform's stored data
Page Creator	Create and host web pages	work-related and personal interests
Phone book	United States street address and phone number information	personal and business contacts
Picassa	Edit and share photos	possible exposure of entire computing platform's stored data
Reader	Web-based feed reader	work-related and personal interests
Ride Finder	Find ground taxi using real time position of vehicles	current location, home address, work address
Scholar	Scholarly paper search	research interests, current school
Search by number	Numeric searches of databases	patent searches, packages you are expecting
SMS	Text messaging interface to Google services	work-related and personal interests
Spell Checker	Alternative spelling for queries	words you do not know how to spell
Stock quotes	Live stock quotes and information	investment portfolio
Talk	Voice and IM communication	possible exposure of your online communications
Toolbar	Access Google services from browser toolbar	possible exposure of entire computing platform's stored data
Transit	Plan trips using public transportation (Portland, U.S. only)	current location, home address, work address
Translate	View web pages in alternative languages	languages you speak
Travel information	Status of flights in United States	travel plans, preferred airlines, location of home
University Search	Constrains search to specific university	research interests
Video	TV program and Video Search	work-related and personal interests
Weather	Weather conditions for any location United States	current location and travel plans
Web Accelerator	Improve web performance	possible exposure of entire computing platform's stored data
Web Search	General web search	work-related and personal interests
Who links to you	Pages that point to a specific URL	your personal or work web page

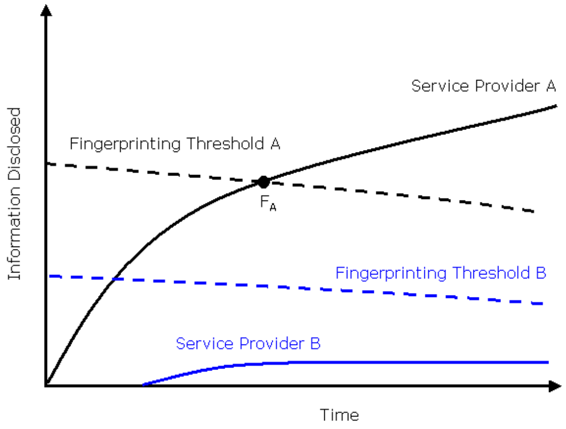

Figure 4: Fingerprinting Over Time: This figure depicts the sum total of information provided to information service providers. Based on the nature of the information disclosed and the resources of the provider there exists a threshold required to uniquely tie this information to an organization or individual.

3. The Fingerprinting Threat

Information that we disclose individually or collectively as organizations is usually less sensitive if it cannot be uniquely identified or fingerprinted. We believe that this mapping is possible in almost all instances given enough interactions with the information service provider. Figure 4 depicts this relationship. In this example a user or organization discloses information to two information service providers over a period of time, as shown by the Service Provider A and B curves. Each information service provider possesses some capability to fingerprint activity given enough information, as shown by the lines Fingerprinting Threshold A and B. The slope of these two lines is negative because we assume that over time the two information service providers will acquire improved fingerprinting capabilities, perhaps due to advances in data mining technology. If, at some point, the information disclosed to an information service provider exceeds the fingerprinting threshold, the user or organization will be uniquely identified. This occurs with Information Service Provider A at point $F_{A,}$ but never occurs with Information Service Provider B as the information disclosed never exceeds their fingerprinting threshold. Based upon unique characteristics of the interaction it is possible, at times, to immediately identify a user even when using a different computing platform.

While there are many ways to attempt to fingerprint users and organizations we address five: network addressing, cookies, browser environment variables, registered user accounts and content/behavior based fingerprinting.

Network Addressing - In order to access resources from an information service provider on today's Internet, the connection will almost certainly occur via the IP and TCP protocols. Assuming this TCP over IP exchange, the user will need a valid IP address to communicate.[§§] At a minimum, the information consumer, in order to communicate, will divulge basic information in the header fields of IP and TCP packets. While some of this information may be obfuscated, fields such as source IP address are difficult to spoof and still allow communication to take place. We believe that most end users and organizations currently make little effort to obscure IP addresses beyond those provided by occasional dynamic allocation and network address translation at organizational firewalls. In many instances the IP address can be used as a unique, or nearly unique key to assist fingerprinting.

Cookies - Cookies are designed to fingerprint users and are supported by all common browsers. Cookies are issued by information service providers and are passed to, and stored by, web browsers. Unless the user has configured their browser to the contrary, the cookie is presented to the information service provider. It is trivial for an information service provider to issue a cookie containing a unique key to identify the browser as well as compel users to enable cookies in order to use their service. While cookie management tools are available, we believe they are not widely used and most users accept the promiscuous default behavior of their browsers.

Browser Environment Variables - As part of the HTTP exchange, browsers pass environment variable data that includes the browser's host operating system, browser type and the user's preferred language. While this information can be easily spoofed, we believe most users do not adjust their browser's default settings to do so. Even if spoofed, the user may inadvertently create a unique set of environment variables that would assist in quick fingerprinting. In the great majority of cases, browser environment variables are not sufficient to fingerprint the user, but can be combined with other sources of information to speed fingerprinting.

Registered User Accounts - User accounts are a requirement for many web-based services. Unless shared, the username/password combination uniquely identifies the user.

Content/Behavior Based Fingerprinting - Content and behavior-based fingerprinting presents a serious threat. While there are countermeasures described in the next section that address cookies, network addressing and environment variables, the actual content of information disclosures combined with nuances of user behavior is more difficult to mask. As an example, consider a user who checks a distinct portfolio of stocks on a daily basis using Google's stock quote service.

The range of potential fingerprinting techniques discussed above represent some ways to uniquely tie together disparate user and organizational data. In our opinion, they are just the beginning and we leave a more complete list for future work. In the next section we describe countermeasures to help reduce the effectiveness of these techniques.

[§§] Although, in rare instances, TCP hijacking is a possibility.

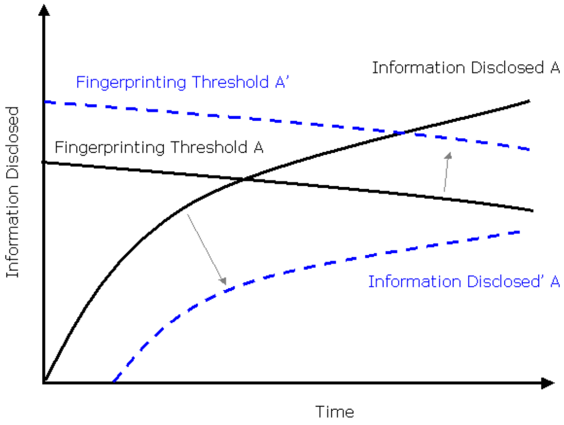

Figure 5: The Impact of Privacy Countermeasures on Fingerprinting. The use of privacy enhancing countermeasures will raise the fingerprinting threshold and reduce the amount of information disclosed. To maintain anonymity, the information disclosed should not cross the fingerprinting threshold.

4. Existing Countermeasures

Information that we disclose individually, or collectively as organizations, provides the possibility of fingerprinting our activities. Countermeasures seek to disrupt our online signature and reduce the likelihood that such fingerprinting will occur. More specifically, these countermeasures increase the time required to perform fingerprinting by reducing the type and quantity of information that we disclose. If properly executed, countermeasures will deny certain key elements required for fingerprinting and increase the adversary's fingerprinting threshold. Consider Figure 5. In this example, the user or organization initially discloses information to Information Service Provider A as seen by the curve labeled Information Disclosed A. Information Service Provider A possesses the resources to fingerprint the user when the information disclosed crosses Fingerprinting Threshold A. By applying countermeasures, the information disclosed is reduced as seen in the curve Information Disclosed A'. Ideally, this new curve will not cross the Fingerprinting Threshold for the duration of their online activities with the information service provider, but the actual deltas depend upon the combined impact of the employed countermeasures. In addition, if the countermeasures are effective at denying information critical to fingerprinting, such as IP addresses, the threshold may shift upward to Fingerprinting Threshold A' and provide an additional measure of safety.

Countermeasures vary widely in the amount of protection they provide, their usability and the degree of adoption by users. The following describes categories of existing countermeasures and our assessment of three metrics: protection level, adoption level and usability. We consider these categories and assessments as initial estimates, for future work we plan to conduct user and organizational studies to validate and refine these results.

Diverse Online Personas (Protection Level: Low, Adoption Level: Low, Usability: Medium) - This category includes countermeasures that distribute data disclosure across multiple online accounts (including email addresses and e-commerce accounts) and information service providers. Protection is minimal due to the small number of information service providers and the management overhead required to maintain multiple accounts with a range of information service providers.

Diverse Computing Platforms and Network Connectivity (Protection Level: Low, Adoption Level: Medium, Usability: Medium-High) - By using a variety of computing platforms and network connection paths, it is possible to increase diffuse online activity. As an example, a user may use multiple computing devices such as a desktop, laptop, personal digital assistant (PDA) or cell phone when using online services. Usability in this case would be reasonable if this was their normal way of conducting online activity. As another example, Internet service providers will often assign dial-up*** customers a temporary IP address each time a user connects to the Internet. Again usability would be high, as the user would not be required to take any additional actions beyond their normal behavior. An ambitious user might even employ virtual machines in an attempt to display alternate operating systems and browsers. Ultimately, we believe each of these measures will be largely ineffective as it is difficult to utilize or simulate a large number of computing platforms and network addresses over an extended period. Given the relatively small number of possibilities, likely less than 20, clusters of activity can still be tied together using the fingerprinting techniques we described in the previous section.

Network and User Aggregation (Protection Level: Low (NAT) High (Crowds), Adoption Level: High (NAT) Low (Crowds), Usability: High (NAT) Low (Crowds)) - The inverse of diverse computing platforms, network aggregation countermeasures seek to tie together a number of network users in a way that makes it difficult to isolate the activities of a single user. This may occur consciously through techniques such as ATT's Crowds system or inadvertently through Network Address Translation (NAT) firewalls.

Network Anonymization Proxies (Protection Level: High, Adoption Level: Low, Usability: Medium-High) - Proxies mask network address information and make web browsing appear to originate from a number of locations. Proxies bear the greatest promise in providing anonymity to network users. Usability can be quite high. In some instances, such as Anonymizer.com, the user need only browse the Web via the Anonymizer website. Other proxies are available, such as the SwitchProxy [48] plug-in for the Firefox browser, but it requires more user skill than that of simply browsing from a website. In both cases, the adoption level is low. We attribute this primarily to lack of awareness and, in some instances, to the moderate complexity of configuring some proxies. A shortcoming of these countermeasures is the requirement to trust intermediary proxies. Even if fingerprinting is frustrated by the use of proxies and the third parties involved prove trustworthy, the disclosed information, such as web

*** While dial-up service usually uses dynamic IP address allocation, which affords some protection, dial-up service is rapidly losing popularity when compared to high speed Internet connections.

searches, may still provide enough information to allow fingerprinting.

Individual and Organizational Awareness (Protection Level: Medium, Adoption Level: Low, Usability: Medium) **-** We believe that awareness of the threat of inadvertent information disclosure is another critical countermeasure. Users and organization will not give a wholehearted effort to employ countermeasures, even if mandated by their employer, without believing there is a problem. This category includes an alert workforce that is prepared for the threat posed by information disclosure even with trusted service providers. Information security awareness training is active in some organizations, but we believe that the training needs to be expanded to cover information service providers.

Direct Connectivity (Protection Level: Low, Adoption Level: Medium, Usability: Low) - This straightforward countermeasure avoids the use of an intermediary search engine by connecting directly to an information service provider or web site. It is awkward to use as it requires the user to know the domain they wish to visit from memory and affords only minimal protection because interactions with the destination website still occur.

Policy and Legal (Protection Level: Variable, Adoption Level: Variable, Usability: Variable) - Policy and legal countermeasures may run the gamut of strict mandates that protect privacy to ineffective guidelines. They may be employed by governments, information service providers and organizations. Examples include, a law that mandates reporting of inadvertent information disclosures by companies conducting business in a given state, an internal privacy policy at a web search company that dictates the destruction of sensitive information after a given period of time and a strict policy by an employer to restrict external Internet access. Policy and legal countermeasures affect both the end user, information consuming organizations and information service providers. This countermeasure can be extremely effective if users and organizations work with information service providers and law makers to craft policies and laws that protect their privacy, while at the same time, allow business to occur.

Cookie Management (Protection Level: Low, Adoption Level Low, Usability: Medium) - Cookie management tools, such as that in the Firefox browser [49], allow users far greater control of the privacy risk posed by cookies.

Cryptography (Protection Level: Medium, Adoption Level: Low, Usability: Low) - Cryptography is a potent technique to provide confidentiality in many areas of security, but currently offers only limited assistance when applied to the information disclosure problems we are addressing. Information service providers are trusted parties in our communications so typical web security mechanisms, such as SSL and IPsec, do not provide adequate protection. In some instances, such as email, contents of individual messages may be secured, but external information including source and destination email addresses is left unprotected. Depending on policies and procedures implemented by information service providers, cryptography can be used to protect against insider threats by encrypting sensitive user information stored on internal servers.

Abstinence (Protection Level: High, Adoption Level: Low, Usability: Low) - Finally, we believe that abstinence, choosing not to use information services may be useful in certain instances. Individual users and organizations must weigh the benefit provided by disclosing some information against the risk of disclosure and the value of the service provided. Decisions of this type occur every day, but lack of awareness toward the threats of long-term information disclosure and fingerprinting biases the decision toward extremely risky behavior.

We believe that collectively, existing countermeasures can provide a reasonable degree of protection for both individuals and organizations, but unfortunately, such countermeasures are infrequently utilized, due in large part to limited usability and lack of awareness. We argue that existing countermeasures are largely ineffective when one considers typical activity on the order of weeks and months, possibly years. Even with increased awareness, each layer of protection requires additional user overhead, and, in some cases, expert knowledge to configure and operate. In addition, many countermeasures impact performance in a non-trivial way. For example, in informal discussions with Tor users the vast majority no longer used the service because it was too slow. As we look to the future, countermeasures must be able to provide both usable security and a proven degree of confidence in the protection they provide. Ideally, they would be able to assure, not just protect, anonymity. In addition to this theoretical level of assurance, they must also be extremely usable. A key component of this usability is transparency. Our notion of transparency requires silent but assured protection for the user or organization that will likely require anonymization to be tightly coupled to networked applications, operating systems and, possibly, network infrastructure. We provide more specifics on the future work required to achieve these goals in the next section.

5. Future Work

Despite their short existence, it is hard to imagine a future without the tools Google and other large information service providers provide. Virtually every Internet user takes advantage of these tools on a daily basis. But a larger problem *does* exist and today's laissez-faire paradigm must be challenged in order for dialog to occur and to develop solutions that allow these powerful tools to co-exist alongside our personal and organizational privacy. To facilitate these solutions, we propose the following future work.

A critical first step is raising the awareness of companies and individual Internet users. To this end, we believe that tools must be constructed that allow both groups to monitor their information disclosure. Initially, we recommend the construction of a Firefox browser plug-in [50] that provides an easy-to-use, but insightful catalog of the information we disclose, such as search terms. Second, we recommend continued research into *usable* anonymous web surfing. Current solutions do not provide adequate usability, which has greatly hindered widespread adoption. To be most effective, this functionality must be cleanly integrated into the operating system and not via an aftermarket add-on. Ultimately, existing anonymous browsing techniques only provide a partial solution as they primarily address network level identification and offer limited protection against content-level information disclosure. Therefore, we recommend not just anonymous browsing, but high-quality online persona management. If properly executed, persona management would allow users to portray a wide variety of online personas from anonymous to very sensitive, based upon the requirements of their interaction with a given service. The goal of this research would be to give users and enterprises more control of their information disclosure as well as provide tools for monitoring their own activities. Third, we suggest that information service providers use SSL for virtually all web transactions in order to thwart eavesdroppers. Finally, we believe debate should begin regarding potential oversight of Google and similar large information service providers. Laws and policies need to be crafted that enforce rigorous requirements for protecting our information and destroying it after a reasonable period. This oversight will almost certainly become a requirement, if information service providers are unable or unwilling to police their own activities. As part of this discussion, we believe national level notification laws, similar to that enacted by the State of California [51], need to be put in place that require businesses to inform users when information disclosures occur as well as to provide increased transparency.

6. Conclusions

We do not believe that the world would be a better place if companies like Google did not exist. Nor do we believe, in the words of Former Sun Microsytems CEO Scott McNealy, that "privacy is dead, get over it." The essential question is one of balance between maximizing good and minimizing harm. We are disclosing information at a prodigious rate to a handful of large information service providers. This trend shows no signs of abating and effectively provides these select companies a virtual monopoly on our sensitive information. The current paradigm that our behavior is safe is incorrect. Given our assumption that these service providers are unlikely to discard information, over time we freely provide unprecedented clarity on virtually all aspects of our personal and professional lives as well as the organizations we work for and interact with. As a result, we are severely undermining the privacy required to do our jobs. Unfortunately, due to its dominance, innovation, brain trust and powerful free tools, Google is specifically the biggest threat. Future advances in data mining combined with the general apathy and ignorance exhibited by many Internet users will almost certainly allow our activities to be aggregated, fingerprinted and tied to our real world personas. Like water rising behind a dam, we are facing a significant threat that must be addressed. Now and into the future, if we find that our trust in these information providers is misplaced or when the inevitable information disclosing mishaps occur, we will regret the many times we traded our sensitive information in return for free tools and services.

7. References

[1] InternetWorldStats.com. "Internet Usage Statistics - The Big Picture." 31 December 2005. http://www.internetworldstats.com /stats.htm, accessed 28 March 2006.

[2] Google.com. "Corporate Information: Company Overview." http://www.google.com/intl/en/corporate/index.html, accessed 25 March 2006.

[3] ValleyWag.com. "Cause and effect: Schmidt talks, investors talk back." http://www.valleywag.com/tech/eric-schmidt/cause-and-effect-schmidt-talks-investors-talk-back-154574.php, accessed 26 March 2006.

[4] Google.com. "Google Investor Relations: Google Code of Conduct." http://investor.google.com/conduct.html, accessed 26 March 2006.

[5] Google Blogoscoped. "Google Removes Its Help Entry on Censorship." 27 January 2006. http://blog.outer-court.com/archive/2006-01-27-n30.html, accessed 28 March 2006.

[6] Andrew McLaughlin. "Google in China." Official Google Blog, 27 January 2006. http://googleblog.blogspot.com/2006/01/ google-in-china.html, accessed 28 March 2006.

[7] Elinor Mills. "Google balances privacy, reach." CNET News.com, 14 July 2005. http://news.com.com/ Google+balances+privacy,+reach/2100-1032_3-5787483.html, accessed 28 March 2006.

[8] Carolyn Said. "Google says Cnet went too far in googling." San Francisco Chronical, 9 August 2005. http://sfgate.com/cgi-bin/article.cgi?file=/c/a/2005/08/09/GOOGLE.TMP&type=business, accessed 28 March 2006.

[9] Google.com. "Google Corporate Information: Our Philosophy." http://www.google.com/corporate/tenthings.html, accessed 28 March 2006.

[10] Kristin Kovner. "Beyond Google." Smart Money, May 2006, p. 122.

[11] Danny Sullivan. "Searches Per Day." Search Engine Watch. http://searchenginewatch.com/reports/article.php/2156461, accessed 29 March 2006.

[12] Wikipedia.com. "Google search." http://en.wikipedia.org/wiki/Google_(search_engine), accessed 29 March 2006.

[13] Google.com. "Google Language Tools." http://www.google.com/language_tools?hl=en, accessed 28 March 2006.

[14] Google.com "Google More, More, More." http://www.google.com/intl/en/options/index.html, accessed 28 March 2006.

[15] Google.com, "Google Labs: Google's technology playground." http://labs.google.com/, accessed 29 March 2006.

[16] Google.com. "Google Jobs: The engineer's life at Google." http://www.google.com/support/jobs/bin/static.py?page=about.html, accessed 29 March 2006.

[17] Google.com. "Google Information for Webmasters." http://www.google.com/webmasters/bot.html, accessed 30 March 2006.

[18] Internet Assigned Numbers Authority. "IP Address Services." http://www.iana.org/ipaddress/ip-addresses.htm, accessed 30 March 2006.

[19] MaxMind.com. "Go Ahead - Locate Your Internet Visitors." http://www.maxmind.com/app/ip_locate, accessed 30 March 2006.

[20] John Battelle. "The Web Time Axis." http://www.archive.org/web/web.php, accessed 28 March 2006.

[21] Archive.org. "Internet Archive Wayback Machine." http://www.archive.org/web/web.php, accessed 28 March 2006.

[22] Yair Amit. "XSS vulnerabilities in Google.com." Email posting to the Web Security Mailing List. http://www.webappsec.org/lists/websecurity/archive/2005-12/msg00059.html, accessed 30 March 2006.

[23] Andrew Orlowski. “Phishing with Google Desktop.” The Register, 3 December 2005. http://www.theregister.co.uk/2005/12/03/google_desktop_vuln/, accessed 30 March 2006.

[24] Google.com. “Google Privacy Policy: Changes to this policy.” http://www.google.com/intl/en/privacypolicy.html, accessed 29 March 2006.

[25] Howard Mintz. “Feds after Google data.” Mercury News, 19 January 2006. http://www.mercurynews.com/mld/mercurynews/news/13657303.htm accessed 30 March 2006.

[26] Charlie Kaufman, Radia Perlman and Mike Speciner. Network Security: Private Communication in a Public World. Prentice Hall, 2002.

[27] Eric Cole and Sandra Ring. “Insider Threat.” Syngress Press, 2005.

[28] Federal Bureau of Investigation. “Robert Philip Hanssen Espionage Case.” Press Release, 20 February 2001. http://www.fbi.gov/libref/historic/famcases/hanssen/hanssen.htm, accessed 29 March 2006.

[29] Federal Bureau of Investigation. “Famous Cases: Aldrich Hazen Ames.” http://www.fbi.gov/libref/historic/famcases/ames/ames.htm, last accessed 29 March 2006.

[30] ChoicePoint.com. “Privacy at ChoicePoint: 2006 Disclosures of U.S. Data Incidents.” http://www.privacyatchoicepoint.com/, accessed 29 March 2006.

[31] Michael Miller. “Google’s Schmidt Clears the Air.” PC Magazine online, 17 March 2006. http://news.yahoo.com/s/zd/20060317/tc_zd/173753, accessed 30 March 2006.

[32] Slashdot.org. “Beware Your Online Presence.” Slashdot, 19 March 2006. http://yro.slashdot.org/yro/06/03/19/2231257.shtml, accessed 30 March 2006.

[33] Seth Finkelstein. “Google Censorship - How it Works.” http://sethf.com/anticensorware/general/google-censorship.php, accessed 30 March 2006.

[34] EFF.org. “Chinese New Year: Resolutions for Google.” Electronic Frontier Foundation, 30 January 2006. http://www.eff.org/deeplinks/archives/004362.php, accessed 30 March 2006.

[35] EFF.org. “Internet Companies Need Code of Conduct in Authoritarian Regimes.” Electronic Frontier Foundation, 15 February 2006. http://www.eff.org/news/archives/2006_02.php, accessed 30 March 2006.

[36] EFF.org. “Google Copies Your Hard Drive - Government Smiles in Anticipation.” Electronic Frontier Foundation, 9 February 2006. http://www.eff.org/news/archives/2006_02.php, accessed 30 March 2006.

[37] EPIC.org. “Justice Department Subpoenas Search Records; Google Resists.” Electronic Privacy Information Center, 27 January 2006. http://www.epic.org/alert/EPIC_Alert_13.02.html, accessed 30 March 2006.

[38] SearchEngineWatch.org. “Google’s Help Center Page About Censorship Back Online With New Text.” Search Engine Watch, 29 January 2006. http://blog.searchenginewatch.com/blog/060129-152723, accessed 30 March 2006.

[39] Roger Dingledine, Nick Mathewson and Paul Syverson. “Tor: The Second-Generation Onion Router.” Proceedings of the 13th USENIX Security Symposium, August 2004.

[40] Anonymizer.com. “Internet Privacy and Security Solutions.” http://www.anonymizer.com/, accessed 12 April 2006.

[41] Michael Reiter and Aviel Rubin. “Anonymity Loves Company: Anonymous Web Transactions with Crowds.” Communications of the ACM, February, 1999.

[42] Radia Perlman. “The Ephemerizer: Making Data Disappear.” Sun Microsystems Technical Report TR-2005-140, February 2005.

[43] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. “The knowledge complexity of interactive proof-systems.” Proceedings of 17th Symposium on the Theory of Computation, Providence, Rhode Island. 1985.

[44] Michael Rabin. “How to exchange secrets by oblivious transfer.” Technical Report TR-81, Aiken Computation Laboratory, Harvard University, 1981.

[45] Ryan MacDonald, “Web-based User Profiling Using Artificial Neural Networks”, Honours Thesis, Acadia University, 2001.

[46] Dave Cartwright. “Customize Your Content With User Profiling.” Web Developer’s Journal,4 April 2004. http://www.webdevelopersjournal.com/articles/user_profiling.html, accessed 12 April 2006.

[47] Proceedings of the 2005 Symposium on Usable Privacy and Security. http://cups.cs.cmu.edu/soups/2005/program.html, accessed 12 April 2006.

[48] Jeremy Gillick. “Firefox Add-ons: SwitchProxy Tool.” https://addons.mozilla.org/extensions/moreinfo.php?id=125, accessed 13 April 2006.

[49] Mozilla.org. “Using the Cookie Manager.” http://www.mozilla.org/projects/security/pki/psm/help_21/using_priv_help.html, accessed 13 April 2006.

[50] Mozilla.org. “Firefox Add-ons.” https://addons.mozilla.org/extensions/?application=firefox, accessed 30 March 2006.

[51] James Brelsford. “California Raises the Bar on Data Security and Privacy.” FindLaw, 20 September 2003. http://library.findlaw.com/2003/Sep/30/133060.html, accessed 30 March 2006.

The views expressed in this paper are those of the author and do not reflect the official policy or position of the United States Military Academy, the Department of the Army, the Department of Defense or the United States Government.