An Honest Man Has Nothing to Fear: User Perceptions on Web-based Information Disclosure

Gregory Conti
Dept of Electrical Engineering and Computer Science
United States Military Academy
West Point, New York
gregory-conti@usma.edu

Edward Sobiesk
Dept of Electrical Engineering and Computer Science
United States Military Academy
West Point, New York
edward.sobiesk@usma.edu

ABSTRACT
In today's era of the global ubiquitous use of free online tools and business models that depend on data retention and customized advertising, we face a growing tension between the privacy concerns of individuals and the financial motivations of organizations. As a critical foundation step to address this problem, we must first understand the attitudes, beliefs, behaviors, and expectations of web users in order to create an environment where user privacy needs are met while still allowing online companies to innovate and provide functionality that users desire. As security and usability professionals we must identify areas where misperceptions exist and seek solutions, either by raising awareness, changing policy, or through technical means. In this paper, we explore these issues and report the results from a survey of 352 college undergraduates and a comparison group of 25 middle aged adults The results were at times surprising and even contradictory to the views held by security professionals. To summarize our findings, the students we surveyed believe that "an honest man has nothing to fear."

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, Human Factors, Legal Aspects.

Keywords
data retention, web search, googling, privacy, anonymization, usable security, information disclosure, anonymity, fingerprinting, AOL, Google, Yahoo!, MSN

1. Background and Motivation
"If you give me six lines written by the hand of the most honest of men, I will find something in them which will hang him." Cardinal Richelieu (1585-1642)

Does web-based information disclosure and data retention really matter to today's web users? We set out to explore this question by surveying 352 college undergraduates regarding their attitudes, behaviors, beliefs, and expectations associated with their use of the web. The results were surprising and sometimes contradictory. On one hand, the students generally felt comfortable with their privacy on the web -- despite the fact that the vast majority admitted to at times searching for things they would not want their parents or future employers to know about. On the other hand, they professed only guarded trust of top web search engines and e-commerce companies and were largely unaware of the duration that data is retained. They admitted to not knowing how to surf anonymously, but did not seem concerned. To summarize our findings, they spoke with a common voice that their trust of the web is calculated and that they view privacy as their responsibility. With this foundation, we believe they feel an honest man has nothing to fear.

Web-based data retention is an extremely important, but contentious subject. Information pours into the databases of the largest service providers at unprecedented rates. For example, Google users conduct an estimated 100 million search queries per day. These queries are processed by approximately two dozen globally dispersed data centers. These data centers offer nation state level information processing resources provided by an estimated 450,000 servers with four petabytes of RAM, 200 petabytes of storage and bandwidth of 3 petabits per second [1]. Other web-based service providers, in domains such as email and messaging, generate similar amounts of data. For example, Hotmail has 260 million users and MSN messenger has 240 million users. As discussed in previous works, the sum total of these interactions paint unprecedented, detailed portraits of the personal and professional lives of web users, as well as the companies they work for [2, 3].

In the era of web-based information services and e-commerce, long-term data retention is a reality. In fact, it is a best practice for these service-oriented businesses, but due to the potential consumer backlash it is rarely publicly discussed by them. Business models of the largest web based information service companies depend on data retention for customized advertising and hence large portions of their billion-dollar revenue. Similarly, in the search industry, businesses pursue perfect search and individual customization [4], both of which are probably unachievable without some form of user data retention. Formal accounts of data retention policies by top web search companies are scarce, but anecdotal evidence suggests that every interaction with these companies is scrupulously logged and stored indefinitely. In one of the rare instances where this subject is addressed, Usama Fayad, Chief Data Officer of Yahoo! stated that Yahoo! collects 10 terabytes of user data a day, not including content, email or images. Additionally, he stated that the first and largest data mining challenge is the "ability to capture all of this data reliably, process it, reduce it, and use it to feed the many, many reports and applications." [5] Combining the ever

Symposium On Usable Privacy and Security (SOUPS) 2007, July 18-20, 2007, Pittsburgh, PA, USA.

decreasing cost of long-term storage with business models that depend upon the customization of information, we believe (in direct contradiction with some survey participants) that nothing is ever discarded. This belief, as well as the sensitivity of the aggregated data, is validated by the outcry surrounding AOL's inadvertent disclosure of a large search query dataset in August 2006. We believe, in fact, these databases of user interactions represent some of the most highly prized assets of online companies. Further, we argue that the mere existence of information stockpiles of this magnitude guarantees that the data will be coveted by many. Government agencies, law enforcement organizations, and industrial competitors greedily eye the data as a way to seek competitive advantage. Despite being in the best interest of online companies, and their shareholders, to protect the data, there exist many legal (and illegal) mechanisms to gain access. In this paper we seek to contrast these realities, with the attitudes, behaviors and beliefs of web-based information service users. The key contributions of this paper are detailed insights into the following questions: how do typical users employ free web tools, what is their expectation of privacy, and how do they believe their data will be handled, used, and retained. Given today's legal and business environment, we discuss where we feel these users are right and where their perceptions may be incorrect.

This paper is organized as follows. Section 2 places our contribution in the field of related work. Section 3 describes the construction of the survey as well as the demographics of the respondents. Section 4 presents the results of the survey. Section 5 provides detailed analysis as well as emergent themes from the survey. Finally, Section 6 presents our conclusions and suggested areas for future work.

2. Related Work

The novelty of our work springs from our direct examination of end user perceptions on anonymity and privacy in a post-AOL disclosure world. The study of anonymity and privacy when online is not new. Online privacy and data retention have been growing concerns since the World Wide Web was created 16 years ago [6]. Organizations such as the Electronic Frontier Foundation (EFF) [7] founded in 1990 and the Electronic Privacy Information Center (EPIC) [8] founded in 1994 have sought to defend digital rights and focus public attention on emerging privacy issues. These early efforts proved prescient. As global use of the World Wide Web skyrocketed in the late 1990's, Internet users began providing a tremendous stream of data to Internet service providers and web-based companies such as AOL, Google, Amazon, Ebay, Yahoo! and Microsoft. In 2003, John Battelle defined the import of this phenomenon in his Database of Intentions writings [9, 10]. "Google Hacking" by Johnny Long [11] was the first to comprehensively describe how to extract sensitive security information from web search engines, but does not focus on information gathered from the perspective of the search provider. Later, the paper "Googling Considered Harmful" codified the threats and countermeasures associated with many forms of web activity [2].

At the same time as "Googling Considered Harmful" was going to press, AOL inadvertently released a dataset containing approximately 20 million web searches for 658,000 AOL users [12]. This incident brought the issues of web-based information disclosure and data retention, albeit briefly, to the forefront of public debate. Widely covered by the media, the disclosure spawned a spurt of analysis of the incident and its implications. Most notably, New York Times reporters Michael Barbaro and Tom Zeller demonstrated the trivial nature of working backwards from an "anonymous" cluster of web searches in the dataset to the real-world user who created them [13]. In the months following the disclosure, a number of websites were created to provide an easy to use interface to the data and eventually these sites added additional analysis functionality that allowed Internet users to collaboratively examine, rate, and, in some cases, identify each user [14,15,16].

Given the backdrop of the AOL disclosure, our main focus is to determine user perceptions regarding information disclosed to free online services in the context of a perceived "private" interaction. In other words, when the user provides data to the online service, such as a search query, the user believes it should be kept private. It is important to note that we are not addressing interactions where the user expects the information will be published, as in the case of social networking sites [17,18]. Nor are we considering the information disclosed through a suspect host or network, perhaps due to spyware, phishing, an untrustworthy operating system, bootstrapping sequence, or malicious ISP.

Surveys that cover web-based information disclosure and data retention are sparse. In 2002, the Danish Presidency distributed a questionnaire to European Union member nations that covered current data retention laws and mandatory data retention at the nation state level. However, the results have not been released citing security concerns [19]. In 2005, Deloitte conducted a survey of Chief Security Officers in the financial sector that included privacy and data retention coverage, but only from the organizational perspective, not that of typical end users. One important finding from this survey is that only 68% of the respondents in 2005 had a program in place for managing privacy compliance within their organizations and only 25% allowed customers to manager their privacy preferences [20]. Conducted during late 2006, ISP-Planet surveyed Managed Security Providers, again focusing at the corporate level and not individual users, and found that nine providers offered services which monitored web content transferred via HTTP and HTTPS [21]. In 2006 Cisco commissioned a survey studying remote worker security. The survey found that a large percentage of remote workers engaged in "risky" online behavior regarding their work PC's, including online shopping (40%), sharing computers (21%) and opening unknown email messages (38%) [22]. While this study examined user attitudes and behaviors, the key distinction is their focus on a wide range of risky online behavior with minimal emphasis on web-based information disclosure and a study group that examined only remote workers.

Two surveys covered broad user level beliefs on trust of online companies. The first was commissioned by TRUSTe, a nonprofit organization which certifies websites based upon online privacy and email policies. It focused on how users determine the trust of websites and the countermeasures users employ to protect their privacy. It found that the majority of online consumers do not ever read privacy statements provided by websites and that only 20% say they read the privacy statement "most of the time." While their focus was not on web-based information disclosure, they did report important related findings. These findings included the following. Thirty-three percent of survey participants believe they did not provide websites with information that would identify them personally, and 86% of American Internet users believe they know how to protect their

personal information online. In addition, 57% of respondents claim to consistently take the necessary steps to do so [23]. The second study was conducted by researchers at the University of Pennsylvania's Annenberg School for Communication. The survey, conducted before the AOL search query disclosure, studied user knowledge of online marketplace rules and focused on marketing and pricing practices. In addition, 63% of their respondents were age 35 or older. They also reported findings that complement our work. These included a finding that only 17% agree with the statement that "what companies know about me won't hurt me" and 65% say they "know what I have to do to protect myself from being taken advantage of by sellers on the web." [24]

It is important to note that there have been a number of initiatives that have attempted to provide anonymity and privacy when using the World Wide Web. World-wide, legal measures have both mandated and limited data retention by web companies and Internet Service Providers (ISPs). Similarly, there have been a number of technical approaches to online anonymity, including anonymous proxy services such as Anonymizer [25] and SpyNOT [26] as well as anonymity networks such as Freenet [27], I2P [28], and TOR [29] that mask network connectivity through overlay networks. There also exist filtering proxies including Privoxy [30] and Proxomitron [31] that can remove privacy damaging web content, such as cookies or third party advertisements. In addition, users have at their disposal a variety of browser-based countermeasures including basic privacy options that wipe locally stored information including history, cookies, saved form data, website passwords, downloaded files and cached web content. More advanced options are also available including Safe History [32] and Safe Cache [33] which defend against browser history-based web privacy attacks as well as third-party cookie managers [34].

3. Survey Design and User Demographics

In this survey we sought to determine user attitudes and behaviors surrounding web-based information disclosure and data retention. More specifically we sought to explore the following areas:

- Amount of search activity as well as search engines used
- Searches on sensitive information
- Trust of search and ecommerce companies
- Personal responsibility for privacy protection
- Familiarity with data retention including awareness of the AOL dataset disclosure
- User understanding of online anonymity
- The user's desired balance between privacy and functionality

All 352 survey participants were members of an upper level undergraduate information technology course. At their institution, this course is mandatory for all students who are not majoring in an ABET accredited degree program. There were 0.28% sophomores (1), 91.76% juniors (323) and 7.96% seniors (28). The average age was 20.8 years old with a Standard Deviation (SD) of 1.03. Respondents were 18.47% female (65) and 81.53% male (287).

The survey itself was a web-based instrument available to only internal campus machines. In order to reduce bias in the survey, we gave only minimal administrative instructions and no information about the content of the survey. We believe that we minimized self selection by soliciting students from a core course, but some degree of self selection will still be evident because these students chose not to major in an ABET accredited engineering program. The impact is that our study participants likely do not possess an advanced information technology or engineering background. While the participants in the survey were college undergraduates, it is important to note that there are some unique characteristics of the population and their day to day environment, in particular they must abide by a strictly enforced honor code and Internet usage policy.

The survey consisted of 25 questions grouped into the following categories: demographics, web usage, search engine privacy, searches on sensitive information, trust of online companies, data retention and anonymity. We carefully sequenced questions from general web usage to those covering anonymity and privacy, in order to minimize the influence of "scare." In addition, we also asked survey participants not to backtrack and change previously answered questions.

4. Survey Results

In this section we present the results of our survey and provide the general reasoning behind the questions. In Section 5 we provide our analysis as well as the emergent response themes and our key findings.

4.1 Web Usage

We first asked a series of questions targeting web usage, with an emphasis on search activity. Our respondents reported conducting web searches, using a search engine, an average of 63.53 times per week (SD=121.7) and that they have been using search engines for an average of 7.72 years (SD=2.42). Based on these results we surmise that our respondents were, in general, experienced long-time users of search.

We asked two questions to identify which search engines participants used and their reasons for the choice. The first question asked users to select the search engine that they used the most, see Table 1. We allowed only a single response and provided options for four of the most popular search engines in the United States. We also included an "other" option. The second question asked why they used this engine and offered a variety of reasons to choose from, see Table 2. While most of the questions in Table 2 are self explanatory, it is important to note that we deliberately chose the wording "because it came with my computer," to gain insight into their use of search toolbars integrated into browsers, despite the fact that, technically, search engines are websites and are often not "included" with a computer. We also included the question "I use other services from this company" to help detect instances where users may be disclosing information to a single company via non-search activities.

Table 1. Search Engine Popularity

Question	Google	Yahoo	MSN	AOL	Other
Which specific search engine do you use the most?	92.44%	6.4%	0.58%	0.00%	0.58%

Table 2. Reasons for Choice of Search Engine

Question	strongly disagree	disagree	agree	strongly agree
It came with my computer.	29.45%	36.44%	27.11%	7.00%
I feel it provides the best search.	3.78%	6.98%	55.23%	34.01%
It appears to be the most popular.	9.38%	20.82%	51.91%	17.89%
I use other services from this company.	19.30%	49.12%	23.98%	7.60%
It's easy to use.	3.21%	0.87%	37.90%	58.02%

4.2 Search Engine Privacy

The next group of questions sought to determine user attitudes surrounding search engine privacy. After the survey's initial demographic and web usage questions, we asked participants how comfortable they were with the privacy they have when using search engines, see Table 3. As mentioned earlier, we carefully sequenced questions from very general to potentially biasing. To determine our success at minimizing bias, we asked this exact same question at the end of the survey. Based on the minimal difference between responses, we believe that respondents were not unduly influenced by the questions and that we limited skewing of the results.

We also sought to determine the degree to which participants felt personally responsible for protecting their personal information, see Table 3 (bottom). In addition, one of the key questions in the survey forced users to decide between search quality or search privacy, see Table 4. We deliberately did not include a neutral response in order to force respondents to make a choice between the two potential end states. Note that for this question, we assumed a tension between the options.

Table 3. Search Engine Privacy

Question	strongly disagree	disagree	agree	strongly agree
I am comfortable with the privacy I have when I use search engines. (asked at start of survey)	4.89%	15.23%	70.69%	9.20%
I am comfortable with the privacy I have when I use search engines. (asked at end of survey)	4.08%	18.37%	71.43%	6.12%
It is my responsibility to protect my personal information.	2.62%	8.43%	52.33%	36.63%

Table 4. User Prioritization Between Search and Privacy

Question	perfect search	search ahead of privacy	privacy ahead of search	perfect privacy
If I had to prioritize between perfect search and perfect privacy, I would choose…	17.97%	37.39%	34.78%	9.86%

4.3 Sensitive Search

In this set of questions we sought to determine if survey participants used search engines to search for things they believed to be sensitive. We constructed these questions to address searches they would not want their parents or current and future employers to know about, see Table 5. By doing so, we forced the participants to implicitly self identify their criteria for sensitive search interactions.

Table 5. Sensitive Search Queries

Question	never	once or twice	sometimes	frequently
At some point in my life, I've conducted web searches on topics I wouldn't want my parents to know about..	13.41%	20.99%	55.10%	10.50%
At some point in my life, I've conducted web searches on topics I wouldn't want my current or future employer to know about.	18.31%	30.23%	43.90%	7.56%

We also included two additional questions which we, as researchers, believe to be potentially sensitive: vanity surfing and searches for contact information on friends and coworkers, see Table 6.

Table 6. Vanity and Social Relationship Searches

Question	never	once or twice	sometimes	frequently
I've used a search engine to search for my own name.	5.19%	64.27%	26.80%	3.75%
I use search engines to look up friends/colleagues contact info.	18.21%	31.79%	41.33%	8.67%

4.4 Trust of Online Companies

User trust of online companies is a key component regarding user comfort level with web-based information disclosure. To address this issue we included a multi-part question to gauge user

attitudes in this area. The question asked participants to rate the extent that they trust four popular search companies (Google, Yahoo, AOL and Microsoft) and two leading online businesses (Ebay and Amazon) to protect their personal information, see Table 7.

Table 7. Perceived Trust of Leading Online Companies

To what extent do you trust the following companies to protect your personal information?				
Company	little trust	limited trust	reasonable trust	strong trust
Microsoft	6.96%	19.42%	53.62%	20.00%
eBay	14.83%	28.78%	42.73%	13.66%
Google	11.88%	30.72%	48.12%	9.28%
AOL	17.68%	34.49%	43.77%	4.06%
Yahoo	15.07%	34.78%	46.96%	3.19%
Amazon	10.82%	22.51%	51.75%	14.91%

4.5 Data Retention

In this group of questions we sought to gain insight into users' beliefs and understanding of data retention. We believe that data retention is not foremost in the minds of typical users, so we first asked questions to determine if they understood that data retention occurred at all, see Table 8. Note that we asked questions regarding both retention of search queries *and* retention of click-throughs to see if users perceived a difference between these two different interactions.

Table 8. Data Retention Perceptions

Question	never	sometimes	frequently	always
Search engines retain the keywords I search on.	0.88%	12.02%	41.94%	45.16%
Search engine companies retain the links I click on from their search results page(s)?	1.47%	20.23%	38.42%	39.88%

The follow on question then asked participants to assume that data retention occurred to some degree and to estimate the duration which the search company would retain the information, see Table 9.

Table 9. Perceived Data Retention Duration

Question	hours	days	months	years	decades
If search engine companies retain the keywords I search on, I believe they will be retained for.	2.04%	15.74%	37.90%	29.74%	14.58%

Our next question told users to assume that search queries would be retained forever and asked how this would impact on their search habits, see Table 10. Note that we carefully chose this sequence of questions to first identify initial user data retention perceptions, and then asking the user to assume retention occurs for some period of time, and finally to assume that it occurs forever. Our aim with this sequence was to minimize bias.

Table 10. Impact of Data Retention on Search Habits.

Question	no change	minimal change	somewhat of a change	significantly change
If you knew for a fact that the topics you search for using a search engine were saved forever, would it change you search habits?	29.28%	39.71%	25.51%	5.51%

Our final question in this group was designed to assess user familiarity with the AOL dataset disclosure. This was a seminal event for most computer security professionals, but we wanted to learn how significant an impact it had on a typical user. We decided to measure this by simply asking, several months after the event, whether a user was aware of the incident.

Table 11. AOL Dataset Disclosure Results

Question	no	vaguely	somewhat	very
Are you familiar with the AOL data disclosure of August 2006?	83.58%	7.33%	7.33%	1.76%

4.6 Anonymity

We asked three questions which focused on anonymity. Our intent was to determine if participants believed their web search activity was anonymous as well as discover how many participants felt they had the ability to search anonymously, see Table 12. We complemented these questions, with a question asking whether the respondent had user accounts with four popular online services, see Table 13. We included this question because we believe registering for a user account uniquely identifies a user to the online service company. Note that this question allowed users to select all that applied.

Table 12. Anonymous Web Surfing

Question	strongly disagree	disagree	agree	strongly agree
I believe my use of a web search engine is anonymous.	19.30%	58.77%	19.88%	2.05%
I know how to surf anonymously.	28.07%	57.02%	13.16%	1.75%

Table 13. Percentage of Users with the Online Accounts

Question	Google	Yahoo	MSN	AOL
I have user accounts with (check all that apply)	22.38%	60.47%	26.74%	50.29%

5. Emergent Themes and Analysis

As we reviewed the survey results, several themes emerged from the data. In the following subsections we present our analysis in terms of these themes.

5.1 The AOL Disclosure did not Occur for the Typical User

Less then 2% of respondents stated that they were very familiar with the incident, and over 83% of the respondents had no familiarity with the August 2006 AOL search dataset disclosure on 658,000 AOL users. Since this data disclosure was arguably the largest online service leak thus far, and one which received broad media attention, we were surprised by the extreme lack of awareness by our respondents.

To identify if this lack of awareness was limited to only our undergraduate sample, we conducted a second much smaller study involving 25 non-technical middle aged adults. The average age for this smaller survey was 40 years old (SD=7). Our intent with this second group was to survey typical adults from non-technical disciplines. The results of this survey, which are also shown on Figure 1, remarkably matched those of undergraduates. In our second survey, 0% of respondents stated that they were very familiar with the incident, and 84% of the respondents had no familiarity with the incident. From this we conclude that for the typical user (at any age level) the AOL data disclosure essentially did not occur. It appears that the incident did little to raise public awareness toward the issue of web-based information disclosure.

As one reviews the remaining analysis, it should be kept in mind that the vast preponderance of respondents were not aware or biased by the AOL disclosure. Fortunately, none of our other survey questions assumed any familiarity with the event, but rather we progressively built up the respondents' assumptions, using carefully sequenced questions, to the reality of the event and then asked them how they would react.

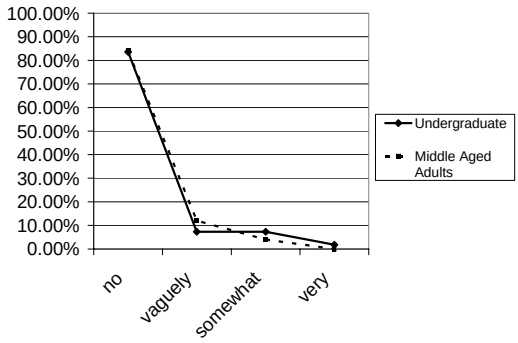

Figure 1. Responses to the question "Are you familiar with the AOL data disclosure of August 2006?"

5.2 An Honest Man Has Nothing to Fear

User trust is calculated. A variety of factors lead us to conclude that users are both comfortable and calculating in their trust of on-line services. Our questions revealed a somewhat mature undergraduate user who:

- states that they are comfortable with their online privacy,
- who demonstrates that they are comfortable with their privacy by conducting sensitive searches,
- who does not feel the need to seek anonymity for their online activities,
- who vigorously use web services but do not blindly trust the companies that provide them,
- who, while not completely understanding data retention, replied that they would only minimally alter their online behavior when asked to assume complete data retention,
- and who ultimately views privacy on the web as a personal responsibility.

Each of these factors, discussed below, portray a user who likes, believes in, and to a measured extent, trusts web services.

About 80% of users agreed that they are comfortable with the privacy they have when they use search engines, see Table 3. They also implicitly demonstrated that they are comfortable with their privacy because 86% of respondents admitted to having at least once searched for topics they would not want their parents to know about and 81% admitted to having at least once searched for topics they would not want their employers to know about, see Table 5. 94% of users have conducted vanity searches on themselves and 81% have looked up contact information on friends and colleagues, see Table 6. In the area of anonymity, only about 22% saw their search activities as anonymous. Only 15% claimed to know how to search anonymously. This all adds up to users who understand that they are at times disclosing sensitive information but are still comfortable with their online privacy. Figure 2 visualizes the responses regarding the four categories of sensitive search.

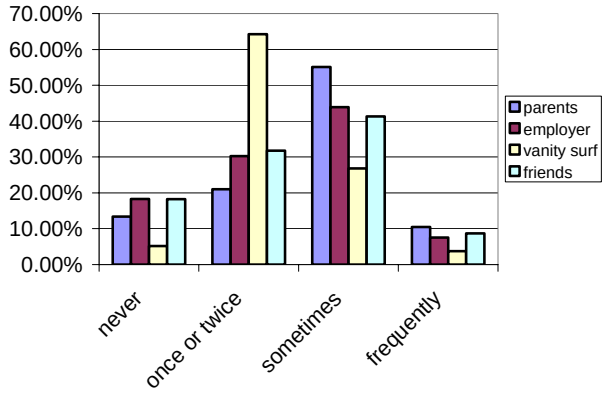

Figure 2. Summary of responses to sensitive search questions including searching for information respondents would not want parents or employers to know about, vanity surfing and searching for friends.

Users also demonstrate their privacy comfort level by prolific use of web services. With an average number of 63.5 searches per week, users clearly are not timid about disclosing information in return for free services (in this case, search). Users also were fairly at ease disclosing additional personal information (giving up a further degree of anonymity) in order to gain additional products and services by registering for online accounts. Table 13 showed that large percentages of respondents possess online accounts. The company with the least number of accounts was Google (with only 22% of the respondents possessing a Google account). 26% of respondents had an MSN account, 50% had and AOL account, and 60% possessed a Yahoo account. As far as number of accounts per user, Figure 3 shows how many accounts a respondent possessed (from the four companies). Note, that over 82% of respondents have at least one online account.

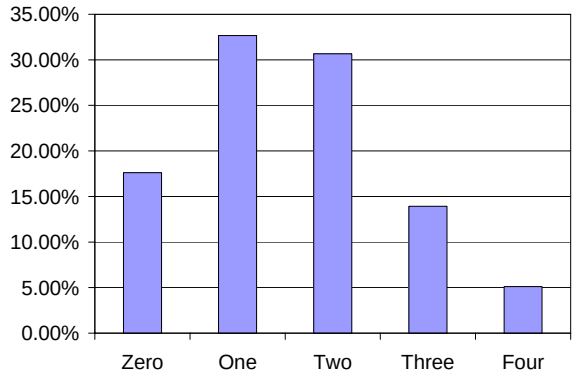

Figure 3. Total number of online accounts possessed by each respondent.

Although respondents use online services and even have accounts with many online companies, they clearly do not blindly trust the companies, see Table 7. Figure 4 graphically depicts the percentage of trust that users have regarding their personal information with the companies of Microsoft, eBay, Google, AOL, Yahoo, and Amazon. The vast majority of users have limited-to-reasonable trust of the online companies. When one looks at the middle two categories of trust (limited and reasonable), the total aggregated percentages are: Microsoft – 73%, eBay – 72%, Google – 79%, AOL – 78%, Yahoo – 82%, and Amazon – 74%. The large percentages for these middle two categories of trust indicate to us that users basically calculate the cost benefit when dealing with online companies and do not strongly distrust or trust the companies.

In the area of data retention, our respondents realize that retention occurs in some degree; they were not far off our assumption of indefinite duration; and the vast majority of respondents indicated that they would not significantly change their search habits even if their searches were retained forever.

99% of respondents believed that at least some of the search keywords are retained and 87% felt that the retention occurred frequently or always, see Table 8. We were somewhat surprised at this result since such a large percentage of respondents were unaware of the AOL data disclosure.

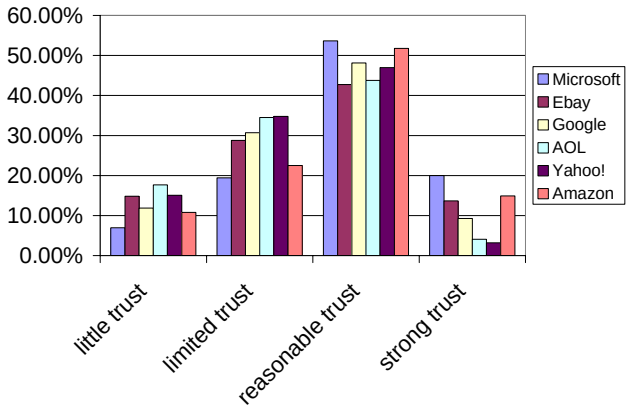

Figure 4. Trust of Major Online Companies. Note that most respondents feel a measured amount of trust.

Respondent estimates of keyword retention duration were not far off the estimates of industry analysts who generally believe that total and permanent data retention occurs in the domain of search. Only 18% of respondents believed search keywords were retained for days or hours. The remaining 82% believed search keywords were retained for months, years, or decades, see Table 9.

A particularly interesting result from the survey involved responses to a question that assumed search keywords were retained forever. 94% of respondents indicated that infinite data retention would not significantly change their search habits, see Table 10. Considering that 99% of respondents felt that at least some of their searches were being retained, we infer that these users are already searching under some sort of assumption of data retention.

The finding that users would not significantly change their search habits based on infinite data retention is supported by respondents' view that it is an individual responsibility. 89% of respondents agreed that protection of personal information is their own responsibility.

5.3 A Cultural Shift May Be Occurring

Although not the central focus of our research, we did find several significant differences between our 352 undergraduates and our follow up survey of 25 middle aged adults. In particular, we noticed distinct differences in the areas of self identified sensitive search activities, responsibility for protection of personal information, and prioritization between privacy and search quality.

Figures 5 and 6 show the significant differences in the sensitive search areas between our undergraduate population and the middle aged group. Note that 61% of the middle aged group have never conducted a search that they would not want their parents to know about. This is in dramatic contrast to the 13% reported by our undergraduate population. Likewise, the self reported sensitive search that the middle aged group would not their employer to know about is also noticeably different.

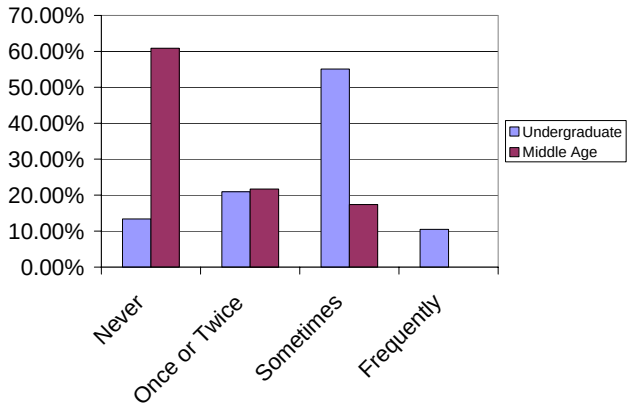

Figure 5. Percentages of undergraduate group and middle aged group who have searched on a topic they would not want their parents to know about.

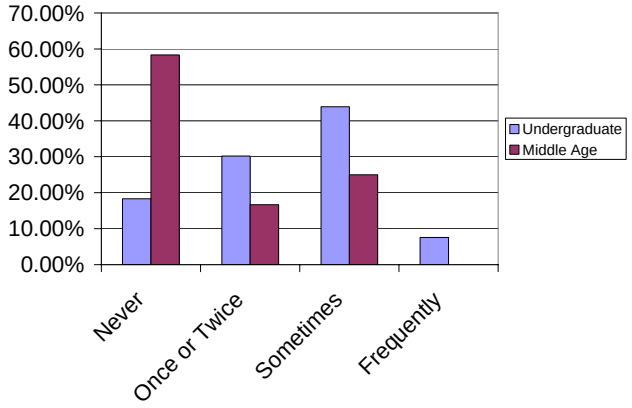

Figure 6. Percentages of undergraduate group and middle aged group who have searched on a topic they would not want their current or future employer to know about.

Additional differences appeared between the two groups in the areas of prioritization between privacy and search quality and responsibility for privacy. When asked to prioritize between privacy and search quality, the undergraduate population placed search ahead of privacy whereas the middle aged group stressed the opposite, see Figure 7. As for responsibility for protecting personal information, almost three times as many undergraduates as middle aged respondents disagreed that it was their personal responsibility, see Figure 8.

While these initial findings suggest some potential differences based on generation, we are hesitant to make any definitive statements and look forward to exploring this issue more in follow-on research.

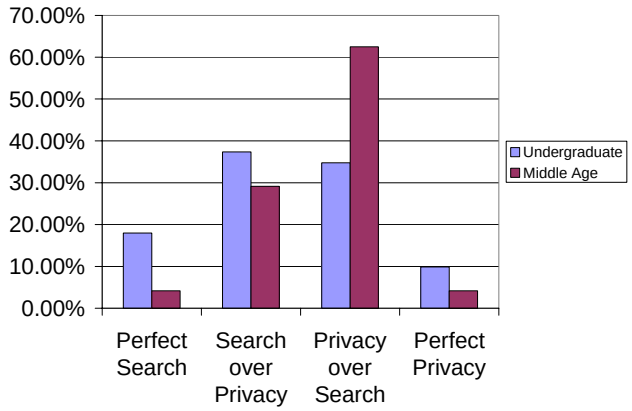

Figure 7. Percentages of undergraduate group and middle aged group who prioritized search and privacy.

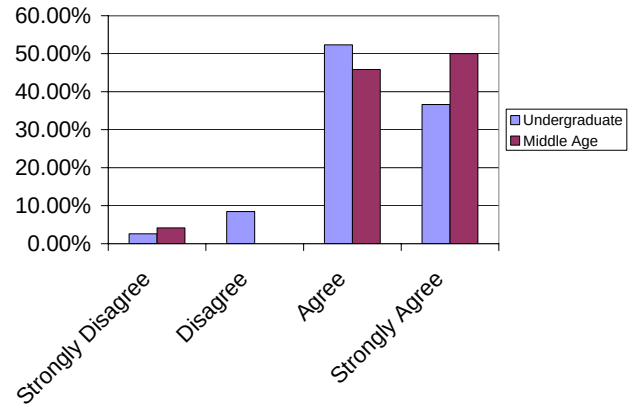

Figure 8. Percentages of undergraduate group and middle aged group who agree that it is a individual responsibility to protect personal information.

6. Conclusions and Future Work

The areas of web based information disclosure and data retention are ripe for future work. Our survey results indicate that a typical user is reasonably aware of data retention and is already working under that assumption. Users admit that they conduct sensitive search activities -- but we feel they underestimate the magnitude of what they disclose when aggregated over time. [1] To address this discrepancy, we believe raising awareness is an important next step. We envision two fruitful initial approaches: integrating coverage of information disclosure and data retention into high school and undergraduate curriculums and creating tools that allow users, as well as organizations, to self-monitor the extent of their information disclosure. The self-monitoring approach will also lead to greater understanding of the magnitude of the problem, both at the individual and organizational levels.

[1] This disparity is also evident in the magnitude of sensitive data deliberatively published online by users of services such as Facebook and Friendster.

In addition to traditional search, we believe there is a need for a comprehensive threat analysis based on *all* web-based data flows as well as research into corresponding self-monitoring solutions. Beyond self-monitoring and raising awareness of the problem of web-based information disclosure, we are concerned about the usability of current anonymous web browsing solutions. Our survey participants were basically unaware of how to surf anonymously. We believe this is because anonymous browsing tools are not widely deployed and anecdotal evidence suggests that these tools are difficult to configure and use by typical users. We suggest future work that seeks to improve usability with the ultimate goal of seamless integration into the web browsing experience.

At the heart of the challenge are business models that seek to improve user experiences and provide targeted advertising by logging and retaining user interactions. These business models should be examined for ways they can be modified to protect anonymity and still provide the incentives and data required by businesses to operate and innovate. For example, how long should data be retained and of what type? There may not be a one size fits all solution. Although some may find it counter to current web culture, there may be the potential for pay-for-anonymity solutions. We believe the most beneficial path lies in collaboration between the large web-based information service providers and their users, not in adversarial relationships, but in cooperative ones that jointly seek effective solutions. Weinstein's "An Open Letter to Google: Concepts for a Google Privacy Initiative" provides one promising roadmap [35].

Our goal in this paper was to study the question: Does web-based information disclosure and data retention matter to web users? To help provide the answer, we surveyed 352 undergraduate students and a comparison group of 25 middle aged adults. In short, we found that users exercise calculated trust of search companies when disclosing sensitive information. In the area of data retention, users seemed largely unconcerned, perhaps because they view protection of personal information as an individual responsibility or possibly because they opt for near-term utility over long-term risk. This balance might shift if survey participants were fully aware of the AOL disclosure, were self monitoring their web-based disclosures, or had the ability to surf anonymously. That being said, users do appear to perform an informal risk analysis as they interact with these companies and services. They trust popular online companies, but only so far. As we look to the future, we expect web users will continue to offer up what appears to be innocuous personal and organizational information in return for free products and services. However, digital information is easily transferred – threats exist and accidents happen. Perhaps an honest man does have something to fear, but today's web users (excluding those violated by the AOL disclosure) don't realize it yet.

7. Acknowledgments

We would like to thank the New Security Paradigms Workshop and Defcon communities for their thoughtful feedback on the problem of web-based information disclosure. We are also grateful to the United States Military Academy's Information Technology and Operations Center for their continued support. In addition, we would like to thank Christa Chewar, Erik Fretheim, Dave Cook, and Jean Blair as well as all the survey participants.

The views expressed in this paper are those of the authors 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.

References

[1] G. Gilder. The Information Factories. Wired, 10.2006, pp. 178-202.

[2] Gregory Conti, "Googling Considered Harmful." New Security Paradigms Workshop, 2006.

[3] Edward Sobiesk and Gregory Conti. "The Cost of Free Web Tools." IEEE Security and Privacy, vol. 5, no. 3, pp. 66-68, May/Jun, 2007.

[4] Dawn Kawamoto. "Google CEO speaks out on future of search." CNET News.com. October 7, 2003. http://news.com.com/2100-1024_3-5088153.html, last accessed 30 January 2007.

[5] Gregory Piatetsky-Shapiro. "Interview with Usama Fayyad, Yahoo Chief Data Officer." ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) Explorations, December 2005. http://www.acm.org/sigs/sigkdd/explorations/issues/7-2-2005-12/fayyad.html, last accessed 6 February 2007.

[6] Robert Cailliau. "A Little History of the World Wide Web." World Wide Web Consortium, circa 1995. http://www.w3.org/History.html, last accessed 16 November 2006.

[7] Electronic Frontier Foundation. http://www.eff.org/, last accessed 6 March 2007.

[8] Electronic Privacy Information Center. http://www.epic.org/, last accessed 6 March 2007.

[9] John Battelle. "The Database of Intentions." Searchblog, 13 November 2003. http://battellemedia.com /archives/000063.php, last accessed 16 November 2006.

[10] John Battelle. *The Search*. Portfolio: New York, 2005.

[11] Johnny Long. "Google Hacking for Penetration Testers." Syngress, 2004.

[12] Ryan Singel. "FAQ: AOL's Search Gaffe and You." Wired News, 11 August 2006. http://www.wired.com/news /politics/privacy/0,71579-0.html, last accessed 6 February 2007.

[13] Michael Barbaro and Tom Zeller. "A Face is Exposed for AOL Searcher No. 4417749." The New York Times, 9 August 2006.

[14] AOL Stalker. http://www.aolstalker.com/, last accessed 6 February 2007.

[15] AOL Search Logs. http://data.aolsearchlogs.com/, last accessed 6 February 2007.

[16] AOLpsycho. http://www.aolpsycho.com/, last accessed 6 February 2007.

[17] Facebook. http://www.facebook.com/, last accessed 13 February 2007.

[18] My Space: A Space for Friends. http://www.myspace.com/, last accessed 13 February 2007.

[19] Statewatch News Online. "EU - Majority of Governments Introducing Data Retention." January 2003.

http://www.statewatch.org/news/2003/jan/12eudatret.htm, last accessed 18 February 2007.

[20] Deloitte. "2005 Global Security Survey." 22 June 2005. available online at http://www.deloitte.com/dtt/research/0,1015,sid=1013&cid=85452,00.html.

[21] ISP-Planet. "Managed Security Provider Survey." 27 January 2007. available online at http://www.enterpriseitplanet.com/security/features/article.php/3656046.

[22] Cisco. "Understanding Remote Worker Security: A Survey of User Awareness vs. Behavior." 2006. Available online at http://www.cisco.com/en/US/netsol/ns340/ns394/ns171/ns413/networking_solutions_white_paper0900aecd8054581d.shtml.

[23] TRUSTe. "TRUSTe/TNS Survey Press Release." 6 December 2006. http://www.truste.org/about/press_release/12_06_06.php, last accessed 24 February 2007.

[24] Joseph Turow, Lauren Feldman, and Kimberly Meltzer. "Open to Exploitation: American Shoppers Online and Offline." Annenberg Public Policy Center, June 2005. Available online at http://www.annenbergpublicpolicycenter.org/04_info_society/Seventeen_Facts_WEB_FINAL.pdf

[25] Anonymizer. http://www.anonymizer.com/, last accessed 13 February 2007.

[26] SpyNOT Anonymous Proxy and Web Browsing Privacy Protection. http://www.spynot.com/, last accessed 13 February 2007.

[27] The Free Network Project: A Distributed Anonymous Information Storage and Retrieval System. http://freenetprotject.org/, last accessed 13 February 2007.

[28] I2P. http://www.i2p.net/, last accessed 13 February 2007.

[29] Tor: Anonymity Online. http://tor.eff.org/, last accessed 13 February 2007.

[30] Privoxy. http://www.privoxy.org/, last accessed 13 February 2007.

[31] The Web Hitchhiker's Guide to Proxomitron. http://www.proxomitron.info/, last accessed 13 February 2007.

[32] Collin Jackson. "Safe History Firefox Add-on." https://addons.mozilla.org/firefox/1502/, last accessed 18 February 2007.

[33] Collin Jackson. "Safe Cache Firefox Add-on." https://addons.mozilla.org/firefox/1474/, last accessed 18 February 2007.

[34] PC World. "Top 5 Cookie Managers." 18 March 2001. http://www.pcworld.com/article/id,44745-page,1/article.html, last accessed 18 February 2007.

[35] Lauren Weinstein. "An Open Letter to Google: Concepts for a Google Privacy Initiative." 9 May 2006. http://www.vortex.com/google-privacy-initiative, last accessed 28 February 2007.