# An Analysis of P3P-Enabled Web Sites among Top-20 Search Results

Serge Egelman
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
egelman@cs.cmu.edu

Lorrie Faith Cranor
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
lorrie@cs.cmu.edu

Abdur Chowdhury
America Online, Inc.
22000 AOL Way
Dulles, VA 20166
cabdur@aol.com

## ABSTRACT

Search engines play an important role in helping users find desired content. With the increasing deployment of computer-readable privacy policies encoded using the standard W3C Platform for Privacy Preferences (P3P) format, search engines also have the potential to help users identify web sites that will respect their privacy needs. We conducted a study of the quantity and quality of P3P-encoded privacy policies associated with top-20 search results from three popular search engines. We built a P3P-enabled search engine and used it to gather statistics on P3P adoption as well as the privacy landscape of the Internet as a whole. This search engine makes use of a privacy policy cache that we designed to facilitate fast searches. Using a list of "typical" search terms taken from AOL users' queries, we examined the trends in privacy policies that are returned from queries to the AOL, Google, and Yahoo! search engines. We then compared these results to results compiled after using "e-commerce search terms" from Google's Froogle service. We examined the top 20 search results returned by each search engine for each of the search terms and found at least one result with a P3P policy for 83% of the typical search terms. Overall we found that these typical search terms yielded P3P adoption rates of 10%. This contrasts with adoption rates of 21% percent when searching for e-commerce terms. Examining the content of the policies, we discovered that a minority of sites engage in direct marketing with or without a way of opting out, and that even fewer sites share personal information with other companies. Finally, we outline ways to increase P3P adoption rates as well as decrease policy errors.

## Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; K.4.4 [**Computers and Society**]: Electronic Commerce



## Keywords

P3P, privacy policies, search engines, e-commerce

## 1. INTRODUCTION

According to a 2005 poll conducted by CBS News and the New York Times, 82% of respondents believed that the right to privacy in the U.S. is either under serious threat or is already lost. This same poll also found that 83% believe that companies may share their personal information inappropriately [7]. These responses are similar to a 2000 survey conducted by The Pew Internet & American Life Project, where 86% of respondents said that they wanted companies to require permission before using personal information for purposes other than those for which it was provided (this includes marketing, sharing with other companies, etc.) [18]. With more recent public scrutiny being paid to media reports of threats to personal privacy ranging from data brokerages to identity theft, there is good reason to believe that more people are becoming concerned with how companies are handling personal data [27]. To address this concern, many web sites are posting their privacy policies for users to analyze. Unfortunately though, most users do not read these policies. A majority of individuals surveyed believed that the mere presence of a privacy policy means that a corporation will not share their data [28]. On the other hand, those who do bother to read privacy policies often cannot understand what the policies say [12]. Thus, privacy policies do not seem to be serving web site visitors well.

The Platform for Privacy Preferences (P3P) was created by the W3C to make it easier for web site visitors to obtain information about sites' privacy policies [10]. P3P specifies a standard XML format for machine-readable privacy policies that can be parsed by a user-agent program. This allows users to specify their privacy preferences to their web browser or other application. When a web site is encountered that does not conform to the user's preferences, the user can be alerted or the agent can take other actions such as blocking cookies.

When a user is trying to locate information on the Internet, more often than not they will use a search engine. Search engines have taken on the role of "gatekeepers of the web" [20]. A January 2005 study found that 84% of all Internet users have used search engines, and an August 2005 study reported that the average user conducts 42 searches each month [16, 5]. Because of the prevalence of search engines in a user's online experience, it would be ideal for a

user to know the privacy policies of all search results without having to visit every site. Most P3P user-agents only show privacy information after a user has started to visit a site. This is a problem for two reasons. First, when the user receives information on how that particular web site will treat their information they have already given them HTTP clickstream information (IP address, browser version, operating system, etc.).[1] Secondly, since the user is already visiting the site, they may be less motivated to visit a different site even after learning about the contents of their privacy policy.

In an attempt to bring privacy information to users earlier in their interaction with web sites, AT&T Labs researchers developed a prototype "privacy-enhanced search engine" that annotates search results with P3P information [11]. When a search term is entered, the search engine retrieves the P3P policies for all of the resulting hits and compares them with one of three levels of privacy preferences.

We extended this work to develop a more robust P3P search service called Privacy Finder. While the AT&T prototype often took thirty seconds or longer to return search results, Privacy Finder typically returns results in less than a second due to our new caching architecture. We have also improved the user interface, adding the ability for users to specify custom privacy preferences and choose between the Yahoo! and Google search engines, and providing links to web site privacy policies and English translations of the XML P3P policy in the search results. Finally, we now reorder the search results so that within each group of ten results those with P3P policies are presented at the top and those matching a user's preferences are presented first.

Our next steps are to evaluate the usability of Privacy Finder and to assess its usefulness in helping web users find web sites that meet their privacy needs. Regardless of user interface design, if Privacy Finder searches rarely return P3P-enabled web sites in the search results or if most of the P3P-enabled sites it finds have privacy policies that users find unacceptable, the tool will not be all that useful to users. Therefore, the present study was designed to assess the extent to which Privacy Finder is able to find P3P-enabled web sites and the level of privacy protection these sites offer. Our study also provides insights into the overall P3P adoption trends three-and-a-half years after P3P 1.0 became a W3C Recommendation. We will first give a detailed description of P3P and the tools used in conducting this study, as well as background information on a 2003 study that looked at P3P adoption rates. Next, we will explain the methodology behind this study. We will then examine the results in terms of comparisons between the different search APIs, overall P3P adoption rates, the types of policies found, and errors in the policies found. We will use our results to offer some insights into how to increase P3P adoption rates. Finally, we present some ideas for future studies.

# 2. BACKGROUND

## 2.1 P3P 1.0

[1]Of course users already provide this information to their search engine as well. Privacy Finder has a policy of only storing logs for one week before deleting them, but other search engines have widely varying policies. Users should always check the privacy policy for each search engine that they use. This all comes down to a question of trust.

The Platform for Privacy Preferences (P3P1.0) Recommendation was issued by the W3C in April of 2002. It has been implemented in the Microsoft Internet Explorer 6 and Netscape Navigator 7 web browsers. P3P specifies an XML syntax for privacy policies, a protocol for user-agents to locate P3P policies on web sites, and a syntax for compact policies sent in HTTP response headers. There are three ways of retrieving a P3P policy:

1. The well-known location: `/w3c/p3p.xml`
2. As part of the HTTP response headers
3. A `<link>` tag in the HTML

Of these methods, the well-known location is the most popular and easiest to implement (77.2% of the P3P-enabled sites we visited for this study use the well-known location). However, it requires access to a particular directory on the web server, which isn't an option for some web site operators.

AT&T Labs researchers developed a P3P user-agent, Privacy Bird, which works with Microsoft Internet Explorer and allows users to specify privacy preferences. When a site is encountered that conflicts with the specified preferences, a red bird is displayed (with an optional audio alert) to notify the user. Conversely, when a site is encountered that complies with the user's preferences, the bird turns green. Implementing such a user-agent relies on completing a series of tasks, the first of which is trying to locate a P3P policy on a target web site. If a policy exists and can be located, it is evaluated against the user's preferences. The preferences are stored in an "APPEL" file. APPEL stands for A P3P Preference Exchange Language, and is a W3C working draft [9]. The language is based on XML and allows users to write one or more rules that specify how an individual's data is to be treated. Once a P3P policy and an APPEL ruleset are entered into an evaluator, a response is returned indicating whether or not the policy conflicts with the stated preferences. At this point, Privacy Bird alerts the user by displaying a red or green bird. Other P3P user-agents may take other actions such as blocking access to the site or blocking cookies from the site.

The W3C runs a P3P validation service that can be used to check the syntax of P3P policies and to make sure P3P files have been setup properly on a web site. The Perl code for this is made freely available [22].

The P3P standard was created to increase understanding of web site privacy policies. However, it is not without its critics. Some claim that industry pushes for self-regulation prevent the U.S. from passing a comprehensive privacy law and leave users with far weaker alternatives [21]. Still others claim that the standard is hard to implement, lacks enforcement provisions, and will never have enough adopters for it to gain momentum [13]. While some of these are valid concerns, we believe that the standard needs to be examined within the context of the current privacy policy environment. A P3P policy is as legally valid as its natural language counterpart. In this paper we address the issue of adoption and do not cover these other concerns as other literature has sufficiently touched on these issues [25].

## 2.2 The Privacy Finder Service

| | Number in list | Sites reached in 2003 | P3P-enabled in 2003 | Sites reached in 2005 | P3P-enabled in 2005 | Percent change |
|---|---|---|---|---|---|---|
| PFF Random | 302 | 286 | 12.23% | 282 | 10.99% | -10.14% |
| PFF Most Popular | 85 | 84 | 30.95% | 84 | 25.00% | -19.22% |
| PFF Refined Random | 209 | 195 | 14.87% | 195 | 12.82% | -13.79% |
| Key Measures | 500 | 486 | 23.46% | 474 | 23.63% | +0.72% |
| Netscore Top 500 | 500 | 488 | 22.95% | 474 | 23.84% | +3.88% |
| Alexa | 500 | 495 | 18.59% | 470 | 18.51% | -0.43% |
| FirstGov | 344 | 338 | 2.07% | 321 | 32.40% | +1465.22% |
| Froogle | 1017 | 1010 | 13.17% | 964 | 12.55% | -4.71% |
| News | 2429 | 2398 | 9.42% | 2286 | 13.56% | +43.95% |
| Yahooligans! | 900 | 868 | 3.00% | 841 | 6.18% | +106.00% |
| **Total** | **5856** | **5739** | **10.25%** | **5414** | **13.59%** | **+32.59%** |

**Table 1: Revisiting the 2003 study on P3P-adoption.**

The Privacy Finder service is largely implemented using a series of Perl scripts. These are served via our Apache server which is running mod_perl. Mod_perl creates a Perl interpreter within Apache so that our scripts are persistent, thus saving time by not having to load an interpreter with each hit. Once a user enters a search term and selects a set of privacy preferences, the selected search API is used to obtain a list of ten search results. The Google API is accessed via the SOAP protocol, while the Yahoo! API is accessed with REST (both protocols are XML-based and run over HTTP). For every search result returned, the web site is contacted in an attempt to locate a P3P policy using all three of the aforementioned methods.

Once a policy is found, it is evaluated against the user's stated preferences. This is done through a stand-alone P3P evaluator engine that is based on Privacy Bird. Finally, the results are reordered and displayed to the user.

One of the biggest problems with trying to retrieve P3P policies from every search result was the obvious performance lag. To address this issue, a large policy cache was created. The P3P specification requires that policies remain valid for a period of no less than 24-hours [10]. Thus, if a policy is already in the cache, there is no need to retrieve it again for 24-hours. Furthermore, when a policy does expire, retrieving it only when a user requests it incurs a burden on the user by forcing him or her to wait longer to see the search results. With these considerations, we created a back-end script which updates the cache every 24 hours. This greatly improved the speed with which a user sees his or her search results. This optimization also facilitated our ability to conduct a study that required running tens of thousands of searches.

## 2.3 Previous Work

In the summer of 2003, the first automated study of P3P adoption was conducted [6]. This study checked for P3P policies on ten lists of URLs. Three of these lists came from the Progress and Freedom Foundation, which had conducted a study in 2001 of corporate web site privacy policies. These lists consisted of popular web sites, a random sampling of web sites, and a refined list of random web sites [2]. One of the lists that was used came from the July 2002 comScore Media Metrix netScore Standard Traffic Measurement report, and contained the top 500 domains with the most unique visitors. This list was used in two previous studies on P3P adoption that were conducted by Ernst & Young [14, 15]. Another list used was the comScore Media Metrix Key Measures, another top 500 list that also included third parties such as advertisers. Another list contained the top 500 domains from the Alexa Traffic Ranking as of February 2003.

The last four lists were created by the researchers after crawling various sites. Froogle was used to create a list of 1,017 commerce-related sites [19]. Yahooligans!, a web index run by Yahoo! and geared towards children of ages 7-12, was used to create a list containing 900 sites. Firstgov was crawled to create a list of 344 U.S. government web sites. Finally, Google News was crawled to create a list of 2,429 news-reporting sites. In total, 5,856 unique sites were examined, 588 of which were P3P-enabled. In addition to comparing our search engine data with this data, we also re-examined the lists of sites used in this previous study. Our findings can be seen in Table 1.

Of the 5,856 unique sites examined, 5,739 were accessible in 2003, and 5,414 were accessible when we repeated this study in February of 2006. The results here show that overall there was an increase in total P3P adoption over the past two and a half years. The total percentage of sites with P3P policies has increased by over 32% as compared to the 2003 study. Additionally, we see very prominent increases in a few small areas. The sharpest increase comes from government web sites. This increase is due largely to the E-Government Act which mandates government agencies post machine-readable privacy policies on their web sites [26]. Additional increases can be seen with regard to news-related sites as well as web sites targeted at children. With regard to the latter, this can also be explained by

legislative initiatives; the Children's Online Privacy Protection Act (COPPA) took effect on April 21, 2000 [29]. It mandates, among other things, that sites targeted towards children under 13 must prominently display a privacy policy which explains how information is to be collected and used. We assume that not every site became compliant under the act immediately after it took effect, which would explain why the number of sites that use P3P nearly doubled from 2003 to 2006. Additionally, this increase isn't as prominent as the one seen on government web sites because COPPA, unlike the E-Government Act, does not mandate that sites post machine-readable privacy policies (the fact that some do is merely a side effect of a larger increase in the number of sites that post privacy policies). While the biggest increase was due to government adoption of P3P, we can also see that the news-related web sites increased P3P adoption by almost 44%.

While these results show an increase in P3P adoption, it is not clear what impact this has on a web user. The URLs used in the 2003 study were taken from lists of popular sites and from lists of sites for a few different industries. While popular web sites are likely to be viewed by more web users, they don't give us a complete picture of which sites are "typically" visited by web users. Thus, we decided to look at search engine results in hopes of getting a better idea of the types of P3P policies and adoption rates when conducting "typical" searches. Research has shown that most search engine results do not appear on lists of popular web sites [17].

| Category | Number of Terms | % of Total |
|---|---|---|
| Autos | 691 | 3.46% |
| Business | 1,213 | 6.07% |
| Computing | 1,076 | 5.38% |
| Entertainment | 2,520 | 12.60% |
| Games | 475 | 2.38% |
| Health | 1,197 | 5.99% |
| Holidays | 325 | 1.63% |
| Home | 763 | 3.82% |
| Misspellings | 1,305 | 6.53% |
| Organizations | 891 | 4.46% |
| Other | 3,128 | 15.64% |
| Personal Finance | 326 | 1.63% |
| Places | 1,225 | 6.13% |
| Pornography | 1,437 | 7.19% |
| News | 1,170 | 5.85% |
| Research | 1,354 | 6.77% |
| Shopping | 2,041 | 10.21% |
| Sports | 659 | 3.30% |
| Travel | 618 | 3.09% |
| URL | 1,356 | 6.78% |

**Table 2: Category breakdown for AOL users' searches.**

# 3. METHODOLOGY

We obtained a list of 19,999 unique search terms entered by AOL users in 2005. These were taken from actual searches, thus we consider them to be "typical" search queries. The terms were randomly sampled from a complete weekly query log. This particular sample size was used because it provides generalizable statistically significant results. AOL staff members manually classified each term into one or more of the twenty categories shown in Table 2.

| Warn when... | Low | Med | High |
|---|---|---|---|
| ...site collects health or medical info for analysis or marketing. | X | X | X |
| ...site shares health or medical info with others. | X | X | X |
| ...site collects financial info for analysis or marketing. | | | X |
| ...site shares financial info with others. | | X | X |
| ...site may contact me by telephone. | | | X |
| ...site may contact me via other means. | | | X |
| ...site does not allow me to remove myself from marketing lists. | X | X | X |
| ...site uses personally identifiable info to analyze me. | | | X |
| ...site shares personally identifiable info with others. | | X | X |
| ...site does not allow me to see the info collected on me. | | X | X |
| ...site uses non-personally identifiable info to analyze me. | | | X |
| ...site shares non-personally identifiable info with others. | | | X |

**Table 3: Table of privacy preference levels.**

The privacy policies of sites where an individual is required to enter personal information are of most concern to us. While every site will receive information such as an IP address and certain browser information, more concern should be given to sites that collect names, contact information, and billing information. Because of this, e-commerce sites stand out. Although many other categories of sites sometimes collect personal information, e-commerce sites consistently collect this information from shoppers. Thus, we also decided to collect search terms from Google's Froogle service [19]. Froogle displays a list of 25 search terms that were recently used. Since Froogle is designed to show products for sale, these terms generally are going to be indicative of e-commerce. Using another Perl script, we screen-scraped these search terms from Froogle. We collected 940 unique terms in this manner.

For every search term, the first twenty hits were examined and stored in our database during the summer of 2005. We conducted Privacy Finder searches with all of the terms in the AOL and Froogle data sets using both the Google and Yahoo! APIs. We also collected the first twenty hits obtained using AOL's search engine for the terms in the AOL data set. For every search term returned, we checked for the existence of a P3P policy. For the sites that did have policies, we then evaluated them against five APPEL rule sets. APPEL rule sets can be used to evaluate a P3P policy according to a particular set of criteria, as discussed in section 2. The first three rule sets were taken straight from the three pre-defined preference settings in Privacy Finder (which in turn were taken from Privacy Bird). These can be seen in Table 3.

The last two rule sets that were used looked to see if a site engages in any marketing practices (excluding opt-in marketing) using personal information, and if a site shares personal information with third parties (excluding opt-in sharing, sharing with delivery companies, and sharing with companies acting as agents for the web site). Finally, using the W3C's P3P validator, we checked to see how many P3P policies contained errors. We saved all of this information in our database for a total of 1,232,955 annotated search hits.

As a benchmark for this study, we examined the 5,856 URLs used in the 2003 study [6], against our database of search results to develop an understanding of how often high traffic web sites appear in search results. Of our 1,122,643 hits, we found that 331,943 (29.57%) correspond to the 5,856 web sites in this list. This indicates that over seventy percent of the time when a user uses a search engine, they are presented with sites that are not on this list. Therefore, examining search engine results may yield data that is more applicable to the user experience.

We also obtained a list of the 30,000 most clicked on domains from AOL search results collected during October of 2005. This list included the number of clicks made to each domain during that period. We checked each of these domains for P3P policies. Of the 30,000 domains, 2,564 unique domains (8.54%) had P3P policies. However, examining the number of clicks to these sites, we found that these 2,564 domains accounted for 16.67% of the total traffic. This further implies that the more popular a site is, the more likely they are to implement P3P. This trend can be seen in Figure 1.

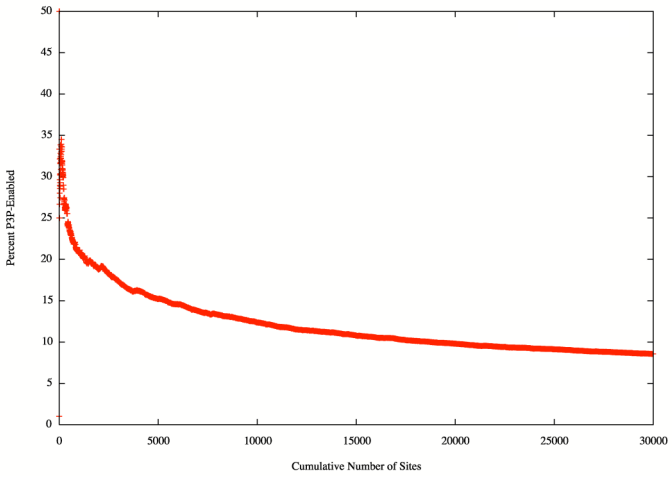


**Figure 1: Plot of web site popularity versus P3P adoption rate. For instance, the 5,000 most popular sites have a P3P adoption rate of roughly 15%.**

# 4. ANALYSIS

We examined the results of our search queries to determine how much choice search engine users have with regard to privacy policies, as well as the rate of P3P adoption as seen through three different search engines. We examined both the quantity of the P3P-enabled hits as well as the "quality" of the policies. Additionally, we examined the shortcomings we uncovered and how to improve by such measures as reducing error rates, increasing P3P adoption across popular sites, and by enforcing legislative measures.

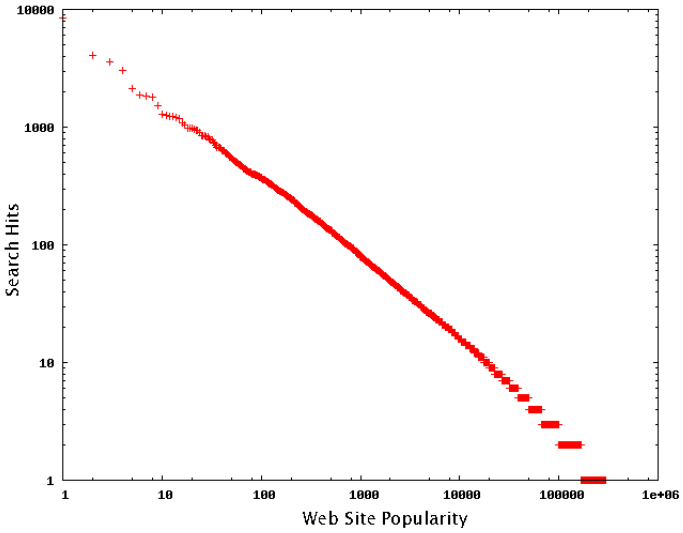


**Figure 2: Plot of web site frequency in search results following a power law.**

In determining the significance of our findings, the data from our results was reformatted so that we could use the SPSS statistical package. Because multiple treatments were used (in this case search engine APIs— Google, Yahoo!, and AOL) to estimate a single factor, we decided to conduct an analysis of variance (ANOVA). Specifically, we conducted four different ANOVAs on our data. First we looked at the AOL data in determining whether the search API used had any effect on the number of P3P-enabled sites that were returned on average with a given search query. This test used Google, Yahoo!, and AOL as the treatments, and the number of P3P-enabled hits as the dependent variable. Next, using the same data set, we wanted to see if the search API played any role in the types of policies returned. That is, would one search API be more or less likely than the others to return better policies. Finally, we repeated these two tests using the Froogle data set.

## 4.1 Overall P3P Adoption

Of the unique terms in the AOL data set, 19,362 yielded search results. This corresponded to 1,160,203 search hits from AOL, Google, and Yahoo!. Of these, 113,880 search results (80,427 were unique) went to URLs that had P3P policies available (10.14%). However, not all of these policies are unique; many of the hits came from multiple different pages on a single domain. In some cases, multiple domain names use the same policy, often because they are owned by the same company. So of the 113,880 P3P-enabled search hits, only 3,846 unique policies were found.

Using the 940 unique search terms from Froogle, 37,560 results were retrieved. Of these, 7,996 had P3P policies, or 21.29%. These correspond to 650 unique policies.

Overall, there are a relatively small number of sites that get returned by the search engines frequently. Specifically, the top twenty most popular P3P-enabled sites account for over 50% of the total number of P3P-enabled hits discovered. The rate at which pages are returned seems to follow a Zipf-like distribution (the frequency trend follows a power law), as shown in Figure 2. This distribution is very similar to

the one depicted in Figure 1, where we examined the 30,000 most popular domains.

Additionally, we also found that many different domains all refer to the same P3P policy. This is largely due to one company owning many different web sites, many companies being owned by one large company (subsidiaries), or because web hosting customers are using the policy of their service provider. In most cases, it is not obvious to the user that this is happening. For instance, while *travel.yahoo.com* and *finance.yahoo.com* both point to Yahoo!'s P3P policy, so do such sites as *geocities.com* and *supermediastore.com*. Largely what this means for the user is that when searching for a particular term, there is a good chance that the P3P-enabled hits will refer to a small number of unique policies, and thus the user has a relatively small number of privacy choices. Though, as mentioned earlier, these larger sites are more likely to have P3P policies than less popular web sites.

The main purpose of having a P3P-enabled search engine is to give users a choice when trying to locate a web site. Rather than simply choosing based on which site has the best title and description as displayed by the search engine, now they can choose based on which site complies with their privacy preferences. Of course, this is not all that useful if few or none of the sites returned have P3P policies available.

As mentioned previously, when a user encounters a site that meets their privacy preferences, a green bird is displayed. A red bird is displayed when the site conflicts with the preferences, and no bird appears when the site does not have a policy. Ideally, a search will yield multiple green birds from which to choose. However, this is often not the case, and in fact it changes based on which search API is being used. These results are shown in Table 4. We found that over 83% of the typical searches included at least one P3P-enabled site in their top twenty results and over 68% of searches included at least one P3P-enabled site in their top ten results. Overall, there was at least one green bird present in the top ten search results on the lowest setting with every search API roughly thirty percent of the time. One notable difference, though, is that Google yielded far more search queries where four or more P3P-enabled sites were listed in a single search; almost twice as many as Yahoo! and AOL.

In addition to the number of P3P-enabled sites available during a given search, we believe that the position of these sites within the search results is also important to the user. While the Privacy Finder service reorders the search results to put the P3P-enabled sites at the top of the results, we examined what positions they tended to be in originally to get some measure of the relevance of the hits to the user's search. Overall, both Yahoo! and AOL tended to show P3P-enabled sites at the beginning of the search results. On the other hand, Google listed these sites in no discernible order; though looking at the first twenty results, those using P3P tended to appear between results eleven and twenty. These distributions can be seen in Figure 3. The effect of ordering is small enough, however, that it is unlikely to be perceived by users.

**Google**

| Hits | Low | Medium | High | Share | Market |
|---|---|---|---|---|---|
| 1 | 31.80% | 26.05% | 17.07% | 30.26% | 28.30% |
| 2 | 14.09% | 10.67% | 5.95% | 13.42% | 11.93% |
| 3 | 7.31% | 5.30% | 2.71% | 7.05% | 5.98% |
| 4 | 4.21% | 2.86% | 1.44% | 3.96% | 3.34% |
| 5 | 2.72% | 1.83% | 0.84% | 2.47% | 2.10% |
| 6 | 1.86% | 1.26% | 0.62% | 1.68% | 1.41% |
| 7 | 1.39% | 0.92% | 0.42% | 1.19% | 1.02% |
| 8 | 0.91% | 0.59% | 0.28% | 0.80% | 0.67% |
| 9 | 0.57% | 0.34% | 0.16% | 0.47% | 0.41% |
| 10 | 0.27% | 0.14% | 0.05% | 0.20% | 0.19% |

**Yahoo!**

| Hits | Low | Medium | High | Share | Market |
|---|---|---|---|---|---|
| 1 | 36.41% | 29.64% | 16.92% | 33.68% | 31.04% |
| 2 | 15.45% | 10.96% | 4.99% | 13.77% | 12.24% |
| 3 | 5.47% | 3.34% | 1.01% | 4.73% | 3.98% |
| 4 | 2.08% | 1.20% | 0.31% | 1.79% | 1.44% |
| 5 | 0.88% | 0.44% | 0.09% | 0.64% | 0.55% |
| 6 | 0.38% | 0.17% | 0.02% | 0.26% | 0.23% |
| 7 | 0.22% | 0.08% | 0.01% | 0.12% | 0.11% |
| 8 | 0.08% | 0.02% | 0.01% | 0.05% | 0.05% |
| 9 | 0.05% | 0.01% | 0.00% | 0.03% | 0.04% |
| 10 | 0.01% | 0.00% | 0.00% | 0.01% | 0.01% |

**AOL**

| Hits | Low | Medium | High | Share | Market |
|---|---|---|---|---|---|
| 1 | 35.24% | 28.85% | 18.38% | 33.58% | 31.37% |
| 2 | 17.33% | 13.44% | 7.52% | 16.57% | 14.94% |
| 3 | 5.53% | 3.70% | 1.22% | 5.35% | 4.27% |
| 4 | 2.19% | 1.42% | 0.48% | 2.18% | 1.61% |
| 5 | 0.84% | 0.48% | 0.13% | 0.77% | 0.58% |
| 6 | 0.31% | 0.16% | 0.05% | 0.25% | 0.22% |
| 7 | 0.16% | 0.06% | 0.03% | 0.09% | 0.09% |
| 8 | 0.07% | 0.03% | 0.01% | 0.04% | 0.04% |
| 9 | 0.03% | 0.01% | 0.00% | 0.01% | 0.02% |
| 10 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

**Table 4: This table shows the cumulative frequency of P3P-enabled search hits. It also shows how often policies complied with each of our five APPEL rule sets. For instance, using Google, 31.80% of the time there was at least one P3P-enabled site listed in the first ten hits that matched our "low" setting.**

## 4.2 Search Engine Comparison

In addition to examining P3P adoption, we were also interested in examining the differences across various search engines. Google functions by examining the number of links to a particular page, the text on those links, and the number of links to those linked pages [4]. AOL uses Google for its

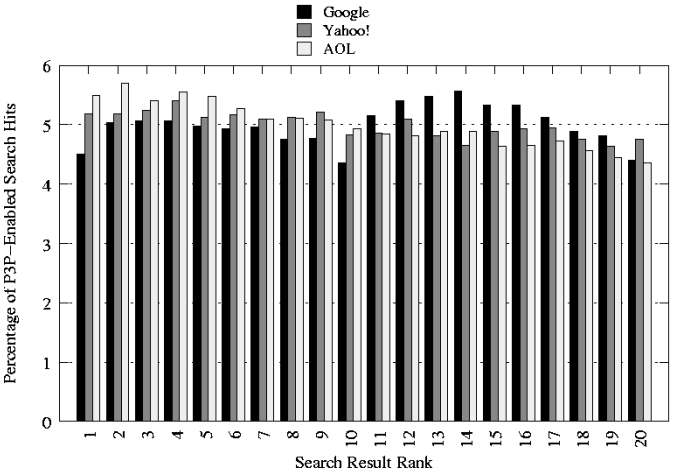


**Figure 3: Distribution of P3P-enabled search results.**

| Search API | Total Hits | P3P-enabled Hits |
|---|---|---|
| Google | 378,183 | 39,574 (10.46%) |
| Yahoo! | 372,819 | 39,055 (10.47%) |
| AOL | 371,641 | 35,251 (9.48%) |

**Table 5: Overview of search API results using the list of "typical" search terms. These results show that Yahoo! yields slightly more P3P-enabled hits than Google, while both yield significantly more than AOL ($p < 0.0005$).**

search service, so we expected largely similar (if not identical) results. Yahoo! on the other hand combines technology from Inktomi, AltaVista, and AllTheWeb. Text matching is done on documents that are found either through spidering, user submission, or paid submissions.

Table 5 depicts the overall rates of P3P adoption across each search API based on the list of "typical" search terms. The number of search terms given to each search API was constant (a total of 19,999 unique terms), but since some terms returned zero hits from one API and a non-zero number from another API, the total number of hits across each API differ. For each comparison, we performed an analysis of variance (ANOVA) with significance set at $p < 0.05$. What is most surprising here is that there is a significant difference between Google and AOL, as AOL uses Google for their searching. We can also see that Google returned slightly more hits than the other search engines— 1.44% more than Yahoo!, and 1.76% more than AOL. Of course, we do not know whether or not these added hits are relevant or which search API returned the most relevant hits overall.

Of all the typical search terms, only 638 of them yielded no results across all three search APIs. This amounts to roughly three percent. We also found that there are a small number of P3P policies that are likely to appear in a large number of search queries. Of these, Yahoo!'s P3P policy is the most prevalent. Overall, there were 31,905 search hits that used this policy, corresponding to 23,335 URLs found on 4,015 different host names. This is because in addition to running a search engine, Yahoo! also offers web hosting services. Thus, when a hosting customer creates a site, they will automatically be using Yahoo!'s P3P policy.[2]

**Typical Search Terms**

| Policy URL | Hits |
|---|---|
| http://privacy.yahoo.com/us/w3c/p3p_us.xml | 31905 |
| http://about.com/w3c/p.xml | 9923 |
| http://privacy.msn.com/p3policy.xml | 3249 |
| http://disney.go.com/corporate/legal/p3p_full.xml | 1688 |
| http://images.rootsweb.com/w3c/policy1.p3p | 1433 |
| http://adserver.ign.com/w3c/p3policy.xml | 1311 |
| http://www.nlm.nih.gov/w3c/policy1.xml | 1159 |
| http://www.bizrate.com/w3c/policy.xml | 1116 |
| http://www.superpages.com/w3c/policy1.xml | 1046 |
| http://www.shopping.com/w3c/statpolicy.xml | 984 |

**Froogle Search Terms**

| Policy URL | Hits |
|---|---|
| http://privacy.yahoo.com/us/w3c/p3p_us.xml | 2320 |
| http://about.com/w3c/p.xml | 590 |
| http://www.bizrate.com/w3c/policy.xml | 562 |
| http://www0.shopping.com/w3c/statpolicy.xml | 212 |
| http://www.shopping.com/w3c/statpolicy.xml | 189 |
| http://www.pricegrabber.com/w3c/p3p.xml | 150 |
| http://www.cpsc.gov/w3c/cpscp3p.xml | 113 |
| http://www.overstock.com/p3p/policy1.xml | 105 |
| http://www.cooking.com/w3c/policy.xml | 94 |
| http://www.altrec.com/w3c/altrec_p3p.xml | 87 |

**Table 6: These tables show the ten most frequently used P3P policies. The first table shows the total hits across all three search APIs (Google, Yahoo!, and AOL) when using the typical search terms, while the second table shows the total hits across the Google and Yahoo! search APIs when using the Froogle search terms.**

Even more interesting though is the number of times Yahoo!'s P3P policy appears when using the Yahoo! search API. While this policy appeared 9,613 (24.29%) times with Google and 9,102 (25.82%) times with AOL, it appears 13,190 (33.77%) times with Yahoo!. This suggests that Yahoo! may give precedence in their search results to their hosting customers. Table 6 shows the top ten P3P policies using both data sets.

## 4.3 Types of Policies

Unfortunately, many users are of the belief that the existence of a privacy policy is indicative of good privacy practices [28]. So while we have shown how various search engines display P3P-enabled sites and how prevalent they are, little information is gained without further examining what sort of practices these policies represent. Table 7 depicts a breakdown of the various policies, listed by the search API and percentage of P3P-enabled sites which resulted in a preference match when evaluated with each of the five rulesets.

At first glance, we can see that one third of all the P3P-enabled sites found do not generate matches at the lowest

[2]We believe that this is actually a problem for Yahoo! and their customers as Yahoo! handles data differently for different hosting customers. Hosting customers who are merchants may or may not use Yahoo! to collect billing information. Additionally, a customer might have privacy practices that are very different than Yahoo!'s.

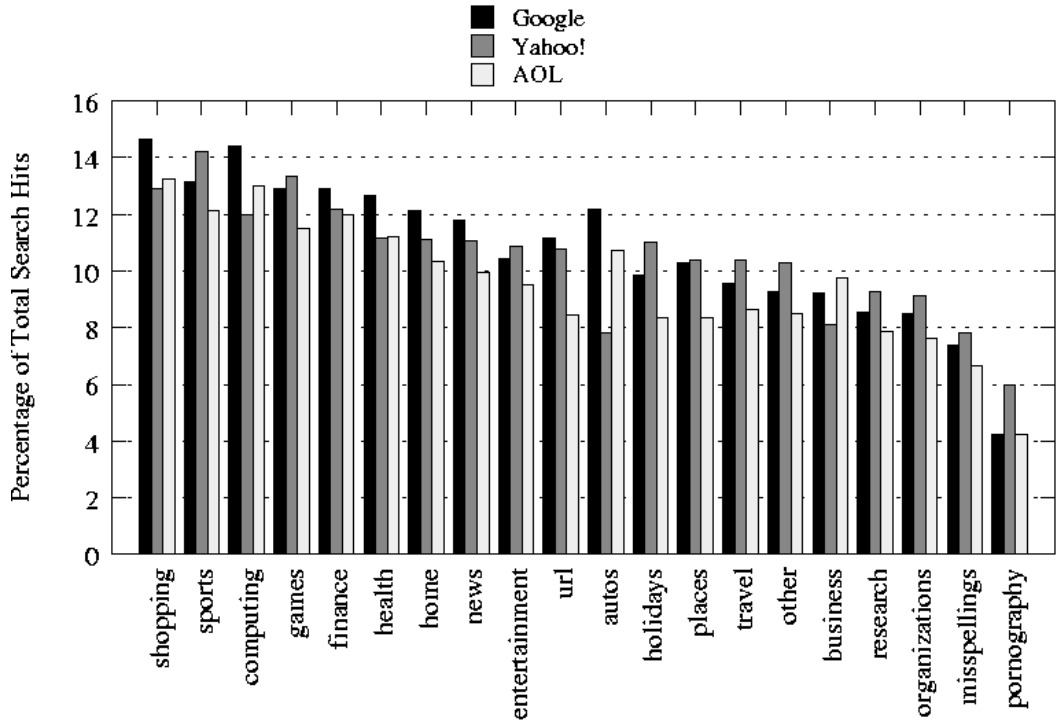


**Figure 4: Distribution of P3P-enabled search results by search term category.**

| API | Low | Medium | High | Share | Market |
|---|---|---|---|---|---|
| Google | 67.65% | 53.47% | 33.23% | 64.33% | 58.21% |
| Yahoo! | 60.35% | 46.81% | 26.18% | 55.17% | 50.27% |
| AOL | 66.85% | 53.46% | 32.02% | 63.77% | 58.01% |

**Table 7: Number of preference matches across search engines using the AOL data. Given all of the P3P-enabled hits returned from a particular search engine, this table shows which percentage complied with each preference level. Google and AOL are statistically more likely to have "better" policies than Yahoo! ($p < 0.0005$), though when compared to each other there is no significant difference in the types of policies that they each return.**

setting. This is because they either collect health information for marketing or sharing purposes, they may contact individuals without providing the option to opt-out, or they do not let individuals remove themselves from their marketing lists. Not surprisingly, two-thirds of all of the sites generate conflicts on the highest privacy setting. Less than half of the sites engage in marketing or sharing. In terms of differences across the search engines themselves, Google and AOL were roughly similar in the types of policies that they returned. Whereas the sites returned by Yahoo! were more likely to conflict with a user's privacy preferences. This may be due in part to the increased likelihood of retrieving sites with the Yahoo! P3P policy while using the Yahoo! search engine. The Yahoo! policy conflicts with all of the preference settings used in this study. As we saw in Table 4, Google is also more likely to have a larger number of hits with green birds from which to choose, thus making it the best choice for the privacy-conscientious user.

Finally, we looked at trends in privacy policies across various industries. The search terms from AOL were accompanied by hand-selected categories. A histogram of the various categories and their rates of P3P-adoption can be seen in Figure 4. While most categories of search terms did not show much differentiation in terms of whether or not they have adopted P3P, there are a handful of observations that can be drawn. Most noticeably, search terms relating to pornography yield sites with significantly fewer P3P policies. Unfortunately, it is difficult to read much into this as there are a number of possible explanations: searching for pornography-related search terms yield fewer corporate web sites, customers may care less about whether pornography-related sites have privacy policies (less demand), etc. Additionally, in terms of aggregate totals, the shopping category yields the most P3P-enabled web sites. This was expected as these sorts of sites are most likely to retain personal information (in order to complete a transaction).

To examine the role of e-commerce web sites further, we compared the results from the AOL search terms to the results found by screen-scraping Froogle for search terms. Overall, the 940 unique terms yielded 37,560 hits using the Google and Yahoo! APIs. Most noticeable was the dramatic increase in P3P adoption— 22.25% of the sites found with Google had P3P policies, and 20.31% of the sites found with Yahoo! had P3P policies. These numbers are roughly 50% higher than what was discovered from using the AOL terms.

Table 8 compares the types of policies found across both search APIs using the Froogle data. This is similar to the data depicted in Table 7; in almost all cases Google yields

| API | Low | Medium | High | Share | Market |
|---|---|---|---|---|---|
| Google | 64.23% | 54.98% | 22.83% | 67.79% | 55.77% |
| Yahoo! | 59.16% | 48.30% | 29.39% | 59.34% | 47.43% |

**Table 8: Number of preference matches across search engines using the Froogle data. Given all of the P3P-enabled hits returned from a particular search engine, this table shows which percentage complied with each preference level. In all cases the differences between the two search engines are significant ($p < 0.0005$).**

"better" policies, since the web sites returned are less likely to share or analyze personal information as well as engage in marketing practices. When comparing Tables 7 and 8 we see that typical searches are more likely than e-commerce searches to return sites that share data with other companies. In addition, typical searches are less likely to return sites that engage in marketing. One hypothesis for this is that it has to do with the reason for collecting information in the first place: the sites returned using the Froogle data are more likely to be commerce web sites that collect data to complete a purchase, whereas the other data set has sites that collect data for different purposes. While some are commerce sites as well, others are sites that may be collecting information as part of a registration form. These types of sites are generally providing free services in exchange for the registration and are thus making money through advertisers, with whom they share this registration data.

## 4.4 Policy Errors

In addition to most sites not implementing P3P policies at all, the majority of sites that we found that do implement P3P had errors in their policies. While a large number of policy errors were noted in the 2003 P3P study, our number is vastly greater [8]. In 2003, one third of the sites discovered contained errors as found by the W3C P3P Validator. However, when using the same validator with our study, we discovered that only about 25% of the total sites examined did not contain any errors. Most of the errors in this study were considered "non-critical errors" in that they conflicted with the P3P specification, but at the same time the evaluator was still able to function correctly. These errors usually amounted to using an older version of the standard. This error can be corrected easily. Additionally, errors were more prevalent across less popular sites. Critical errors, on the other hand, prevented the evaluator from running properly because certain required parts of the policies were either missing or could not be understood (due to syntax errors). The critical errors only accounted for about five percent of all of the URLs found; this number is similar to the 2003 statistic which found critical errors in six percent of the policies examined.

## 4.5 Increasing P3P Adoption

Having a tool that will allow users to make more informed choices about privacy could potentially be very beneficial. Unfortunately, its utility is proportional to the number of sites that choose to publish P3P policies. As we have seen, only around ten percent currently do (from the perspective of the search engine user). This number should be easy to increase, especially as services like Privacy Finder become more popular.

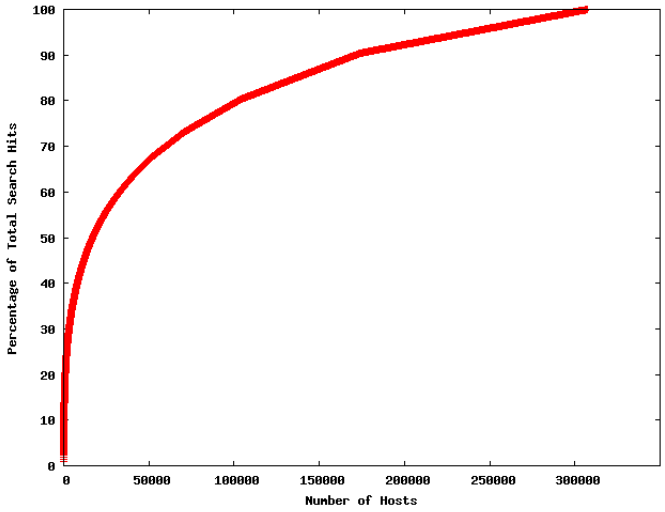


**Figure 5: Cumulative distribution of web sites without P3P policies. The x-axis represents the top $n$ most frequently returned web sites, while the y-axis represents the percentage of total search hits that these top sites account for. For example, the top 100,000 most frequently returned web sites account for 80% of all of the search hits examined.**

In the course of collecting search results, we have noticed that the domain names returned match a power law with regard to the frequency with which they appear in searches. This was previously discussed in section 4.1 and Figure 2. As can be seen here, the most frequently returned web site accounted for roughly 0.7% of the entire collection of web sites returned. This frequency only decreases. Thus, looking at all the sites that do not use P3P, we can see that this follows the same distribution.

The most frequently returned site not using P3P is Amazon.com, which accounts for roughly 0.9% of all of the search hits. Examining the cumulative distribution, we see that the twenty most popular sites that do not use P3P represent 6% of all of our results. Additionally, the top 13,000 account for 50% of our results, as seen in Figure 5. If a small number of these sites were to become P3P compliant, this would create a dramatic increase in the frequency with which P3P-enabled sites get returned by a search engine.

Additionally, the rate of P3P adoption should increase as the result of legislative initiatives. In 2002 the U.S. Congress enacted the E-Government Act. Among other provisions, the act mandates that government agencies publish machine-readable privacy policies on their websites. Since P3P is the only standard for doing this, many government agencies now present P3P policies. The State of Arkansas has since mandated that their agencies follow suit. From our data, we have 24,752 search hits which have ".gov" domain names.[3] Of these, 9,645 (roughly 39%) have P3P poli-

[3]This is just a rough estimate created by searching our cache for domain names ending in ".gov." Some of these domain names belong to state web sites. There are also federal government web sites that do not have a .gov domain name. Thus, we can only make a rough estimate about the rate of government P3P adoption.

cies. On the other hand, examining the ".mil" web sites that were returned, only 173 of the 2,492 queried had P3P policies (6.94%). Combined, this lowers the total rate for government adoption of P3P to roughly 36%. While this is far from being in full compliance with the law, government web sites represent by far the largest sector to adopt P3P.

While a ten percent adoption rate after less than four years might seem paltry, many other W3C standards have taken much longer to gain prominence. For instance, the Cascading Style Sheets 1.0 (CSS) specification became a W3C standard in 1996 [23]. However, it wasn't until four years later in 2000 that any web browser fully supported it (Internet Explorer 5.0 for Macintosh was the first) [24]. Additionally, CSS 2.0 became a W3C standard in 1998, yet as of 2006, there are no web browsers that fully support it [3].

# 5. FUTURE WORK

While this study gave some interesting statistics as to the current rates of P3P adoption, more information can be learned through additional studies encompassing more web sites. Given the number of users using search engines and the different types of information they may be looking for, duplicating this study with a multitude of additional search terms may yield more interesting information, and would certainly yield information about the power and generalizability of this study. For example, it would be interesting to collect search terms from AOL users during different time periods and to collect search terms from other search engines.

Additionally, further examination of the content of the policies might also yield interesting information. As mentioned previously, the majority of the policies found contained errors. Since only about five percent of the errors were critical errors that prevented the execution of the evaluator, most errors can be fixed easily. However, further studies still need to be done to examine how these errors are being introduced, and how responsive webmasters are when asked to fix them. Beyond that, we also plan on looking at semantic errors. Does the language in the P3P policy accurately reflect the natural language policy? To help investigate these questions, we have already written a utility for locating, examining, and archiving P3P policies to examine types of policies as well as how policies change over time. We also hope to learn how representative of the Internet as a whole P3P-enabled sites are.

Finally, to further examine the utility of the Privacy Finder service, we plan on conducting user studies. The main motivation for creating such a service was to facilitate easy understanding and comparison of web site privacy policies without having to first visit the web sites. The success of this project is directly related to how useful such a service is for the users. In determining this, we plan on recruiting users to purchase various items using the Privacy Finder service to locate a suitable merchant. We hope to determine whether or not a company's privacy practices play a role in the user's purchasing decision. Furthermore, previous research has shown that individuals often do not make rational privacy-related decisions, instead opting for small rewards to satisfy immediate gratification [1]. We hope to validate this theory using empirical evidence. This is just one of many possible studies that can be undertaken to determine how useful a privacy-enhanced search engine may be.

# 6. CONCLUSION

With more and more media attention paid to identity theft, data aggregation, and online privacy in general, individuals are starting to care more and more about the privacy policies of the web sites that they visit. Because most users will not take the time to read every privacy policy that they encounter, and many who do are unable to understand the language employed, the P3P specification was created to automate this process. Web browser plugins and other third party applications automatically alert users when they encounter a site that disagrees with their preferences, but this only works after entering the site. The Privacy Finder service was created to solve this problem by allowing search engine results to be annotated with privacy policy information. This allows users to get this information without having to first enter the site.

We have shown that this is a feasible strategy for examining P3P trends on the Internet as most users regularly use search engines. Our data set yielded interesting results regarding the rates of P3P adoption, the types of policies, and differences in various popular search engines. Overall, we have shown that P3P adoption is increasing, both as the result of a public interest in privacy as well as the result of legislative initiatives. While only ten percent of all sites studied use P3P, this number is over twice as high for e-commerce sites. Because most search terms will yield at least one site with a P3P policy, the Privacy Finder service can offer users information to help them make informed choices regarding how they treat their personal information.

# 7. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under grant IGERT 9972762 in CASOS. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS), the Institute for Software Research International, and CyLab at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation or the U.S. government.

The authors would also like to acknowledge AT&T for the development and release of the Privacy Bird source code, on which the code used for this project is based. The previous prototype for the Privacy Finder service was written by Simon Byers, David Kormann, and Patrick McDaniel while at AT&T Labs-Research.

# 8. REFERENCES

[1] A. Acquisti. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the ACM Electronic Commerce Conference (EC 04)*, pages 21–29, New York, NY, 2004. ACM Press. http://www.heinz.cmu.edu/ acquisti/papers/privacy-gratification.pdf.

[2] W. Adkinson, J. Eisenbach, and T. Lenard. Privacy online: A report on the information practices and policies of commercial web sites. Technical report, Progress & Freedom Foundation, 2002. http://www.pff.org/publications/ privacyonlinefinalael.pdf.

[3] B. Bos, H. W. Lie, C. Lilley, and I. Jacobs. Cascading Style Sheets, level 2, CSS2 Specification, May 1998. http://www.w3.org/TR/REC-CSS2/.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th World Wide Web Conference*, 1998. http://www-db.stanford.edu/pub/papers/google.pdf.

[5] E. Burns. Search increased in august, October 7, 2005. http://www.clickz.com/stats/sectors/search_tools/article.php/3554731.

[6] S. Byers, L. F. Cranor, and D. Kormann. Automated Analysis of P3P-Enabled Web Sites. In *Proceedings of the Fifth International Conference on Electronic Commerce (ICEC2003)*, October 1-3, 2003. http://lorrie.cranor.org/pubs/icec03.html.

[7] CBS News. Poll: Privacy rights under attack. *CBS News*, October 2, 2005. http://www.cbsnews.com/stories/2005/09/30/opinion/polls/main894733.shtml.

[8] L. Cranor, S. Byers, and D. Kormann. An Analysis of P3P Deployment on Commercial, Government, and Children's Web Sites as of May 2003. Technical report, AT&T Labs-Research, May 14, 2003.

[9] L. Cranor, M. Langheinrich, and M. Marchiori. A P3P Preference Exchange Language 1.0 (APPEL1.0), April 2002. http://www.w3.org/TR/P3P-preferences/.

[10] L. Cranor, M. Langheinrich, M. Marchiori, M. Presler-Marshall, and J. Reagle. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification, April 2002. http://www.w3.org/TR/P3P/.

[11] L. F. Cranor, S. Byers, D. Kormann, and P. McDaniel. Searching for Privacy: Design and Implementation of a P3P-Enabled Search Engine. In *Proceedings of the 2004 Workshop on Privacy Enhancing Technologies (PET2004)*, May 26-26, 2004.

[12] M. J. Culnan and G. R. Milne. The culnan-milne survey on consumers and online privacy notices, 2001. http://intra.som.umass.edu/georgemilne/pdf_files/culnan-milne.pdf.

[13] Electronic Privacy Information Center (EPIC). Pretty Poor Privacy: An Assessment of P3P and Internet Privacy, June 2000. http://www.epic.org/reports/prettypoorprivacy.html.

[14] Ernst & Young. P3P Dashboard Report. August 2002. http://www.ey.com/global/download.nsf/US/P3P_Dashboard_-__August_2002/$file/P3PDashboardAugust2002.pdf.

[15] Ernst & Young. P3P Dashboard Report. January 2003. http://www.ey.com/global/download.nsf/US/P3P_Dashboard_-_January_2003/$file/E&YP3PDashboardJan2003.pdf.

[16] D. Fellows. Search Engine Users. January 23, 2005. http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf.

[17] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. The egalitarian effect of search engines. Technical report, arXiv.org e-Print Archive, 2005. http://arxiv.org/pdf/cs.CY/0511005.

[18] S. Fox, L. Rainie, J. Horrigan, A. Lenhart, T. Spooner, and C. Carter. Trust and privacy online: Why Americans want to rewrite the rules. August 20, 2000. http://www.pewinternet.org/pdfs/PIP_Trust_Privacy_Report.pdf.

[19] Google, Inc. Froogle, 2005. http://froogle.google.com/.

[20] E. Hargittai. The Changing Online Landscape: From Free-for-All To Commercial Gatekeeping. *Community Practice in the Network Society: Local Actions/Global Interaction*, pages 66–76, 2004. http://www.eszter.com/research/c03-onlinelandscape.html.

[21] H. Hochheiser. The platform for privacy preference as a social protocol: An examination within the u.s. policy context. *ACM Transactions on Internet Technology (TOIT)*, 2(4):276–306, 2002.

[22] Y. Koike and S. Taiki. P3P Validator, January 29, 2002. http://www.w3.org/P3P/validator.html.

[23] H. W. Lie and B. Bos. Cascading Style Sheets, level 1, December 1996. http://www.w3.org/TR/CSS1.

[24] E. Meyer. What Makes CSS So Great?, July 21, 2000. http://www.oreillynet.com/pub/a/network/2000/07/21/magazine/css_intro.html.

[25] D. Mulligan, A. Schwartz, A. Cavoukian, and M. Gurski. P3P and Privacy: An Update for the Privacy Community, March 28, 2000. http://www.cdt.org/privacy/pet/p3pprivacy.shtml.

[26] Office of Management and Budget. About E-GOV, 2005. http://www.whitehouse.gov/omb/egov/g-4-act.html.

[27] H. Taylor. Most People are "Privacy Pragmatists" Who, While Concerned about Privacy, Will Sometimes Trade It Off for Other Benefits. 17, 2003. http://www.harrisinteractive.com/harris_poll/index.asp?PID=365.

[28] J. Turow. Americans and online privacy: The system is broken, 2003. http://www.asc.upenn.edu/usr/jturow/internet-privacy-report/36-page-turow-version-9.pdf.

[29] U.S. Federal Trade Commission. How to comply with the children's online privacy protection rule. http://www.ftc.gov/bcp/conline/pubs/buspubs/coppa.htm.