P3P Deployment on Websites

Lorrie Faith Cranor [a], Serge Egelman [a,*], Steve Sheng [a], Aleecia M. McDonald [a], and Abdur Chowdhury [b]

[a] *Carnegie Mellon University*

[b] *Illinois Institute of Technology*

Abstract

We studied the deployment of computer-readable privacy policies encoded using the standard W3C Platform for Privacy Preferences (P3P) format to inform questions about P3P's usefulness to end users and researchers. We found that P3P adoption is increasing overall and that P3P adoption rates greatly vary across industries. We found that P3P had been deployed on 10% of the sites returned in the top-20 results of typical searches, and on 21% of the sites returned in the top-20 results of e-commerce searches. We examined a set of over 5,000 web sites in both 2003 and 2006 and found that P3P deployment among these sites increased over that time period, although we observed decreases in some sectors. In the Fall of 2007 we observed 470 new P3P policies created over a two month period. We found high rates of syntax errors among P3P policies, but much lower rates of critical errors that prevent a P3P user agent from interpreting them. We also found that most P3P policies have discrepancies with their natural language counterparts. Some of these discrepancies can be attributed to ambiguities, while others cause the two policies to have completely different meanings. Finally, we show that the privacy policies of P3P-enabled popular websites are similar to the privacy policies of popular websites that do not use P3P.

Key words: P3P, Privacy Policies, Search Engines, E-Commerce

[*] Corresponding author.
Email addresses: lorrie@cs.cmu.edu (Lorrie Faith Cranor), egelman@cs.cmu.edu (Serge Egelman), shengx@cmu.edu (Steve Sheng), am40@andrew.cmu.edu (Aleecia M. McDonald), abdur@duvel.ir.iit.edu (Abdur Chowdhury).

Preprint submitted to Elsevier *1 August 2007*

1 Introduction

According to a 2005 poll conducted by CBS News and the New York Times, 82% of Americans believe 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% of Americans are concerned about companies collecting their personal information because of the risk that companies might share their personal information inappropriately [1]. These responses are similar to a 2000 survey conducted by The Pew Internet & American Life Project, in which 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 [2]. To address concerns about their handling of personal data, many websites are posting their privacy policies. However, most users do not read these policies [3]. Furthermore, a majority of individuals surveyed held the mistaken belief that the mere presence of a privacy policy means that a corporation will not share their data [4]. Even those who do bother to read privacy policies often cannot understand what the policies mean [5]. Additionally, websites with poor privacy practices have little incentive to disclose these practices, while websites with good practices may view the posting of their policies as a burden [6]. Thus, privacy policies do not seem to be serving website visitors well.

The Platform for Privacy Preferences (P3P) was created by the World Wide Web Consortium (W3C) to make it easier for website visitors to obtain information about sites' privacy policies [7]. 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 user encounters a website that does not conform to the user's preferences, the agent can alert the user or take other actions such as blocking cookies.

Both end users and researchers may benefit from increasing P3P adoption. P3P best serves end users when a large number of websites with which users share data make their privacy policies available in the P3P format. Even if only a fraction of websites are P3P-enabled, user agents can help users identify the websites that do use P3P, as well as those that have privacy policies that users deem acceptable. Automated tools can also be used to collect and analyze P3P policies for research purposes. This makes it easy for researchers to collect large numbers of policies and compare them across legal jurisdictions or industry sectors, and to track policy changes over time.

This study aims to assess the state of P3P adoption to inform questions about P3P's usefulness to end users and researchers. In Section 2 we provide background on P3P and existing P3P user agents. In Section 3 we present our study methodology. In Section 4 we measure P3P deployment among a number of different sets of websites. In Section 5 we compare the deployment rates we measured with previous studies and present data we collected by monitoring P3P policy additions, deletions, and changes to answer questions about P3P deployment trends. In Section 6 we present our analysis of the content of P3P policies to answer questions about the level of privacy protection offered on the Internet today. In Section 7 we investigate the accuracy of P3P policies to determine how reliable they are and the extent to which they are being kept up to date. In Section 8 we compare the content of P3P policies with the content of human-readable policies at websites that do not have P3P to gain insights into the representativeness of the privacy policies of P3P-enabled websites. Finally, we discuss our conclusions in Section 9. We conclude that while P3P adoption has been slow to date, the number of sites adopting P3P is increasing, and P3P adoption is strongest for e-commerce and U.S. government websites. We show that there are a large number of errors in P3P policies, but most of these errors do not prevent user agents from making accurate assessments of a website's overall privacy level. We also show that P3P policies are generally representative of all website privacy policies and therefore provide a useful data source for website privacy policy studies.

2 The Platform for Privacy Preferences (P3P)

The Platform for Privacy Preferences (P3P1.0) Recommendation [7] was issued by the W3C in April of 2002. It has been implemented in two popular web browsers and in a number of other P3P user agents. The W3C has also issued "notes" describing A P3P Preference Exchange Language (APPEL) [8] and P3P1.1 [9]. APPEL is a language for representing user preferences about P3P policies. P3P1.1 includes a variety of extensions and clarifications to the P3P1.0 Recommendation and documents suggested wording for presenting P3P policy information to end users in English.

P3P was created to increase understanding of website 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 [10]. Others claim that P3P is hard to implement, lacks enforcement provisions, and will never have enough adopters for it to gain momentum [11]. While some valid concerns have been raised, we believe that P3P needs to be examined within the context of the current privacy policy environment in which a P3P policy is as legally valid as its natural language counterpart [12]. In this paper we address the issue of adoption and do not cover these other concerns, which are addressed in other papers [13].

In this section we describe the P3P1.0 Recommendation, some of the P3P user agents currently available, and the Privacy Finder P3P search service we developed.

2.1 P3P 1.0

P3P1.0 specifies an XML syntax for privacy policies, a protocol for user agents to locate P3P policies on websites, and a syntax for compact policies sent in HTTP response headers.

2.1.1 P3P syntax

P3P policies are computer-readable XML documents that provide the name and contact information for the website (`<ENTITY>` element), the types of information that may be collected (`<CATEGORIES>` element), how information may be used (`<PURPOSE>` element), how information may be shared (`<RECIPIENT>` element), information about an individual's ability to access their own information in the site's records (`<ACCESS>` element), data retention policies (`<RETENTION>` element), and options for dispute resolution (`<DISPUTES>` element). A set of multiple choice options are defined for most of these elements, although human-readable fields are also provided to allow for more detailed explanations of privacy practices. In addition, attributes can be used to indicate whether a particular purpose or recipient is always required or whether an opt-in or opt-out policy applies. P3P policies may also contain a `<NON-IDENTIFIABLE>` element if a site does not store personally identifiable data or a `<TEST>` element if the policy has been posted for testing purposes only.

The P3P language is extensible, allowing new elements to be added as needed. These new elements may be labeled as required or optional, indicating whether or not it is safe for a user agent to ignore them if it does not know what they mean.

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 website. The Perl code for validation is freely available [14].

2.1.2 Locating P3P policies

P3P1.0 specifies the format for *policy reference files* that indicate the location of P3P policies on a website and the parts of the website to which they apply. Most websites have just one policy for the entire site; however, some have multiple policies that cover different files or directories on the site. Once a

P3P user agent has obtained a policy reference file, it has the information it needs to locate the relevant P3P policy.

Websites have three options for notifying user agents about the location of their policy reference files. The first option is to place the policy reference file in a standard *well-known location*: `/w3c/p3p.xml`. The second option is to add an HTTP response header that advertises the location of the policy reference file. The third option is to embed an HTML or XHTML `<link>` tag in their HTML content.

The well-known location is the most popular and easiest to implement of these methods (77% 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 is not an option for some website operators.

2.1.3 P3P compact policies

Compact P3P policies consist of a series of tokens transmitted in a P3P HTTP header along with a cookie. The purpose of the compact policy is to enable the web browser to make a quick decision about whether to accept a cookie. The compact policy is only a summary of the site's larger policy, but in many cases is enough for the a user agent to make a decision about a cookie. Every site that uses a compact policy is also required to maintain a full P3P policy so that if more information is needed, the full policy can be analyzed by the user agent. Compact policies consist of a series of three-letter and four-letter tokens separated by spaces. These tokens can represent the multiple choice fields of the following P3P elements: `<ACCESS>`, `<CATEGORIES>`, `<DISPUTES>`, `<NON-IDENTIFIABLE>`, `<PURPOSE>`, `<RECIPIENT>`, `<REMEDIES>`, `<RETENTION>`, and `<TEST>`.

2.2 P3P User Agents

Microsoft's Internet Explorer 6 (IE6) was one of the first P3P user agents available. IE6 allows users to specify personal privacy preferences by selecting from one of the browser's built-in privacy settings or by importing a privacy settings file. These settings are used to specify conditions under which cookies should be blocked or restricted on the basis of their P3P compact policies. IE6 does not consider full P3P policies in its decisions. A small icon is displayed when cookies have been blocked or restricted, but there is no persistent indicator to provide P3P-related information in IE6. IE6 also provides a menu option that allows users to request that a full P3P policy be fetched and displayed in a human-readable format. Informal surveys suggest that very few IE6 users are aware of these P3P-related features.

Netscape Navigator 7 also includes P3P functionality. Much like Internet Explorer, it allows users to choose predefined privacy settings as well as specify custom settings. Again, these preferences apply only to cookies. Netscape also provides a summary of each P3P-enabled site's privacy policy and a link to the site's natural language privacy policy. However, this functionality requires navigating through multiple levels of menus.

AT&T Labs researchers developed a P3P user agent, Privacy Bird, which works with Microsoft Internet Explorer and allows users to specify privacy preferences [15,16]. When users encounter sites that conflict with their specified preferences, the browser displays a red bird in the browser's title bar (with an optional audio alert) to notify the user. Conversely, when users encounter sites that comply with their preferences, the bird turns green. Users can specify their privacy preferences by selecting from pre-packaged "high," "medium," and "low" settings, or by selecting up to 12 conditions to trigger privacy warnings. User preferences are stored in an APPEL file [8], which is evaluated against each site's full P3P policy.

2.3 Privacy Finder

Users frequently use search engines to locate information on the Internet. Search engines have taken on the role of "gatekeepers of the web" [17]. 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 [18,19]. Because of the prevalence of search engines in users' online experiences, it would be ideal for users 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 users receive information on how a particular website will treat their information they have already given the site HTTP clickstream information (IP address, browser version, operating system, etc.). [1] Second, since users are 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 websites, AT&T Labs researchers developed a prototype "privacy-enhanced search engine" that annotates search results with P3P information [20]. 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.

[1] Of course users already provide this information to their search engine as well. Users should always check the privacy policy for each search engine that they use.

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, choose between the Yahoo! and Google search engines, and we provide links to website privacy policies as well as 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.

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, Privacy Finder contacts the website in an attempt to locate a P3P policy using all three of the standard methods. Once Privacy Finder locates a policy, it evaluates the policy against the user's stated preferences using a stand-alone P3P evaluator engine based on Privacy Bird. Finally, Privacy Finder reorders the results and displays them to the user.

Initially, Privacy Finder used the same green and red bird symbols as used by Privacy Bird to indicate a match or mismatch with a user's privacy preferences. Sites without P3P policies received no symbols. However, after user studies showed that users tended to consider sites with red birds worse than sites without P3P policies [21], we replaced the bird symbols with a privacy meter consisting of four square boxes. The four boxes are colored green to indicate a complete match with the user's preferences or white to indicate a complete mismatch. One, two, or three boxes are colored green to indicate partial matches with the user's preferences. When the site is not P3P-enabled, the boxes are not shown at all.

Privacy Finder employs a large policy cache so that users do not have to wait for P3P policies to be retrieved. The P3P specification requires that policies remain valid for a period of no less than 24-hours [7]. 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 that updates the cache every 24 hours.

3 Methodology

This study required the analysis of both machine-readable P3P policies as well as human-readable privacy policies. We adapted Privacy Finder's P3P evaluator as well as the W3C P3P Validator to automate the analysis of the P3P policies. To facilitate semi-automated analysis of the human-readable privacy policies we developed software that displayed each privacy policy along with a data collection interface. We had students read the policies and manually "code" them by answering a series of multiple choice questions about each policy. Our software took the answers to these questions and generated a pseudo-P3P policy to represent each human-readable privacy policy. Because some human-readable privacy policies do not contain all of the information required for a full P3P policy, our pseudo-P3P policies contain elements to indicate points that are "unclear" in the human-readable policy. Each natural language policy took roughly 20 to 40 minutes to code. A second coder verified a random sampling of the coded policies to ensure accuracy.

Once the human-readable policies had been coded into pseudo-P3P policies, they could be analyzed automatically using our P3P tools. We used a set of 67 APPEL files and our P3P evaluator to automatically gather data on what information websites collect, with whom it is shared, whether customers may opt-out of mailing lists, etc. We aggregated the results across various lists of websites, for example to gain insights into trends among the most popular websites versus a random sampling, and across various industries. We also compared the policies of P3P-enabled websites with the privacy policies of sites that do not use P3P. Additionally, we manually coded the natural language privacy policies of P3P-enabled websites so that we could compare them with the P3P policies provided by these websites in order to detect conflicts.

We collected data on privacy policies and P3P policies associated with websites from three sources: Popular and Random site lists we created, the top-20 results from running search queries on three popular search engines, and P3P-enabled websites in the Privacy Finder cache.

3.1 Popular and Random Sites

We obtained a list of the 30,000 most clicked on domains from America Online (AOL) search results collected during October of 2005. This list included the number of clicks made to each domain during that period. We created two data sets for our study based on this list: a "Popular" list and a "Random" list.

The Popular list consisted of the 75 most popular domains on the list. In order to make our results comparable to other studies [22], we compiled this list after removing websites that had a top-level domain other than .com, pornographic websites, and websites targeted to children. Of the 75 websites on our Popular list, 72 had human-readable privacy policies and 21 had both human-readable policies and P3P policies.

We created the Random list in order to study the privacy policies of more "ordinary" websites. We used the top 12,000 most clicked on domains, and then randomly selected one hundred websites, again excluding non-.com domain names, pornography, and children's websites. Of the 100 websites on our Random list, 78 had human-readable privacy policies and 9 had both human-readable policies and P3P policies.

3.2 Search Data

We obtained a list of 19,999 unique search terms randomly sampled from a complete weekly log of search queries entered by AOL users in 2005. We received only the search queries themselves, with no information linking the search queries to the users who entered them or linking multiple search queries together. We consider these search queries to be "typical" search queries. 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 1 [23].

We are most concerned about the privacy policies of sites where an individual is required to enter personal information. While every site will receive information such as an IP address and certain browser information, sites that collect names, contact information, and billing information pose more privacy issues. 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 collected search terms from Google's Froogle service [24]. Froogle displays a list of 25 recently used search terms. Since Froogle is designed to show products for sale, these terms are generally indicative of e-commerce. Using another Perl script, we screen-scraped these search terms from Froogle. We collected 940 unique terms in this manner.

In the summer of 2005 we conducted searches using all of the terms and saved the first twenty results. 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

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 1
Category breakdown for AOL users' searches.

checked for the existence of a P3P policy. For the sites that did have policies, we then evaluated them against five APPEL rule sets. 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.

APPEL rule sets can be used to evaluate a P3P policy according to a particular set of criteria, as discussed in Section 2. We took the first three rule sets 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 2. The last two rule sets were used to check whether 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

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 2
Table of privacy preference levels. The Xs indicate conditions that trigger warnings. For example, if a site collects health or medical information for analysis or marketing purposes, a warning will be displayed for all three preference levels.

sharing, sharing with delivery companies, and sharing with companies acting as agents for the website).

3.3 Privacy Finder Cache Data

Every time Privacy Finder discovers new websites, they are archived in our cache. Thus, we have amassed a collection of sites with and without P3P. Since Privacy Finder is still in beta and regularly undergoes modifications, occasionally the cache is reset. As of December of 2006, Privacy Finder's cache contained over 150,000 websites, 9,408 of which were P3P-enabled. Using Privacy Finder we created another data set of P3P-enabled websites.

While we used the Privacy Finder data set in answering general P3P-adoption questions, we also used it for answering industry-specific questions. Using another script, we screen-scraped the Yahoo! Directory in an attempt to categorize all of these websites. Due to the large number of categories yielded in this fashion, we manually grouped several of them together, and then chose to examine the largest categories: "shopping," "government," "news and media," "computers," "banking and finance," "B2B (business to business)," "adult," "blogs," and "education." In the end we were able to define categories for 16,919 websites (about 11% of the cache). Of these, 1,181 were P3P-enabled (7%).

4 P3P Deployment Rates

We found that the more popular a website, the more likely it is to deploy P3P. We examined the top-20 search results returned by each search engine for each of the AOL 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. We found that Yahoo! and Google yield a similar number of P3P policies, while the AOL search engine yields fewer, despite the fact that it is based on Google. At the same time, we found that Google and AOL yield "better" privacy policies than Yahoo!. We discuss these results in more detail below.

4.1 Overall P3P Deployment

We used the AOL and Froogle data sets to examine P3P deployment in the summer of 2005. 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) were for URLs that had P3P policies available (10.14%). Using the 940 unique search terms from Froogle, we retrieved 37,560 results. Of these, 7,996 had P3P policies, or 21.29%. The difference in adoption rates between the sites found using the AOL and Froogle search terms is similar to the findings of another recent study that found a statistically significant difference in P3P adoption among e-commerce websites versus other popular websites [25]. Yet another recent P3P study found adoption rates of around 25%. This study used a web crawler seeded with popular websites and mostly examined commercial websites [26]. Since their study used a different methodology but arrived at a similar adoption rate for e-commerce websites, we believe our numbers are accurate.

Many P3P policies were counted multiple times in our search hit analysis, as hits often come from multiple pages on a single domain. In addition, multiple domain names sometimes use the same policy, often because they are owned by the same company. The 113,880 P3P-enabled search hits found using the AOL data set correspond to 3,846 unique P3P policies. The 7,996 P3P-enabled search hits found using the Froogle data set correspond to 650 unique policies.

Overall, there are a relatively small number of sites that search engines frequently return. Specifically, the top twenty most popular P3P-enabled domains account for over 50% of the total number of P3P-enabled hits we discovered. The frequency with which search engines return pages seems to follow a Zipf-like distribution (the frequency trend follows a power law), as shown in Figure 1.

When we checked the list of the 30,000 most clicked on domains from AOL search results, we found that 2,564 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 also demonstrates that the more popular a site is, the more likely it is to implement P3P. This trend can be seen in Figure 2.

4.2 Search Engine Comparison

We investigated the differences in frequency of P3P-enabled search results returned by the Google, AOL, and Yahoo! search engines. Google's search result ranking algorithms take into account the number of links to a particular page, the text on those links, and the number of links to those linked pages [27]. AOL uses Google for its 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 3 depicts the overall rates of P3P adoption across each search API based

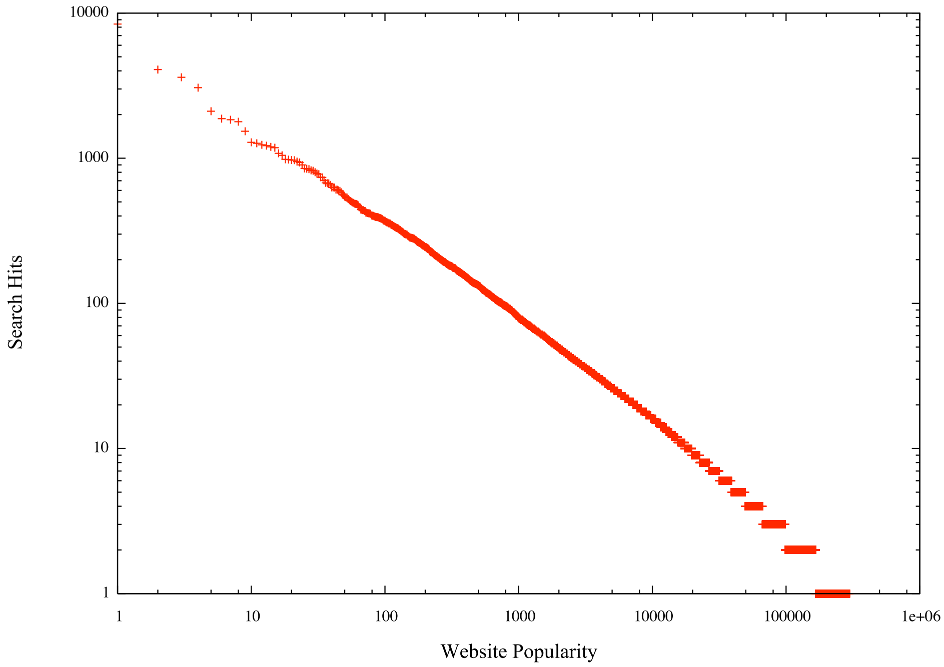

Fig. 1. Plot of website frequency in search results. This data reflects the top 20 search results yielded from each of the 19,999 AOL search terms. The frequency distribution follows a power law.

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 3
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 ($p < 0.0005$).

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 differs. 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, despite the fact that 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.

Only 638 of the typical search terms yielded no results across all three search

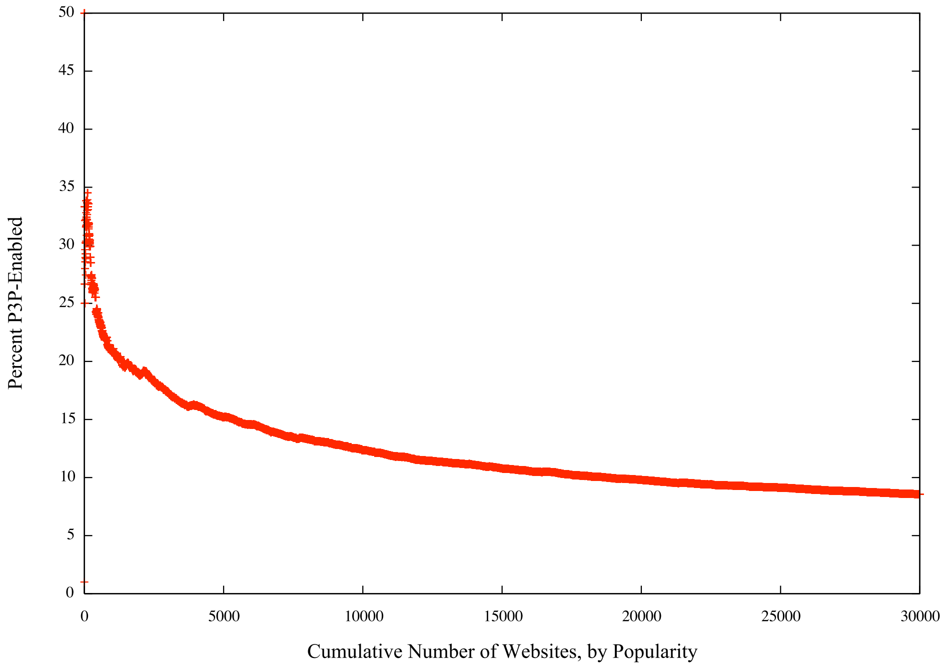

Fig. 2. Plot of website popularity versus P3P adoption rate for the October 2005 sample of 30,000 most clicked on domains. For instance, the 5,000 most popular sites have a P3P adoption rate of roughly 15%.

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 was 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. Yahoo automatically serves their P3P policy at their customer's sites. [2]

Even more interesting 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 4 shows the top ten P3P policies using both data sets.

In addition to the number of P3P-enabled sites returned by 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

[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.

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 4
These tables show the ten most frequently encountered 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.

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. P3P hits are fairly well distributed through the top 20 search results, occuring slightly more frequently at the beginning of the search results returned by Yahoo! and AOL. However, we

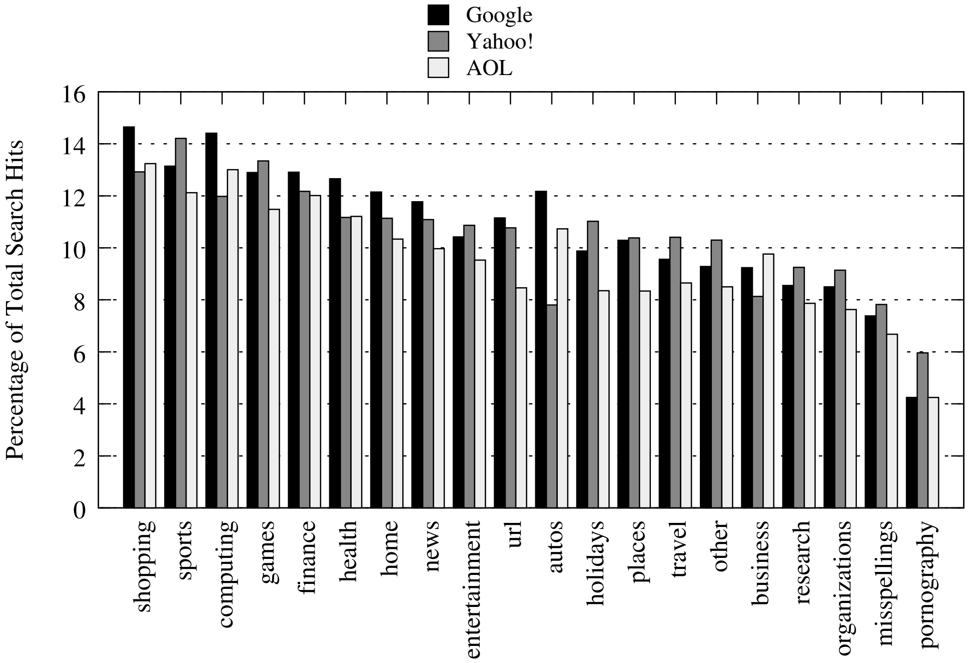

Fig. 3. Distribution of P3P-enabled search results by search term category.

found no statistical significance to this variation. Thus, the effect of ordering is small enough that it is unlikely to be perceived by users.

4.3 P3P Deployment Rates by Website Category

We have seen that deployment of P3P is much higher among e-commerce search results than among the results of typical searches. We used two methods of categorizing P3P-enabled websites to examine P3P adoption rates by website category.

4.3.1 AOL Categories

As discussed in Section 3, the AOL search terms were hand-labeled with 20 categories. We used these categories to explore trends in P3P-deployment rates. Figure 3 shows a histogram of P3P deployment rates across the 20 AOL search term categories. We can see that the shopping category has the highest level of P3P deployment, consistent with our finding that P3P deployment is higher among e-commerce search results than among typical search results. At the other end of the spectrum, search terms relating to pornography yield sites with significantly fewer P3P policies.

4.3.2 Yahoo! Categories

As discussed in Section 3, we used P3P-enabled sites found in Privacy Finder's cache to examine additional P3P trends. We used the Yahoo! Directory[3] to categorize these sites. These categories included 16,919 sites from our cache. Table 5 shows the rate of P3P adoption across the top nine categories. As can be seen, the "shopping," "government," "news and media," and "computers" categories all exceeded ten percent. While the AOL data did not include a category for government websites, the other top three categories show P3P adoption rates that are very similar to the AOL data.

Category	Number of Sites	Number with P3P	Percentage with P3P	Percentage with P3P (AOL Data)
Shopping	2,787	415	14.9%	13.6%
Government	3,050	406	13.3%	N/A
News and Media	1,351	161	11.9%	11.0%
Computers	278	31	11.2%	13.1%
Banking	366	32	8.7%	12.4%
B2B	1,049	73	7.0%	9.0%
Adult	722	20	2.8%	4.8%
Blogs	646	11	1.7%	N/A
Education	6,670	32	0.5%	N/A

Table 5
P3P-enabled sites across categories found using the Yahoo! Directory.

4.4 P3P Around The World

Using Privacy Finder's P3P cache, we discovered 437 P3P policies originating in countries outside of the United States [28]. We used ccTLDs (country code top level domains) to identify the countries of origin. We counted all sites with .com, .mil, .us, .org, .net, .gov, .edu, .info and .biz as US sites. This tends to under estimate the number of non-US sites as websites originating in foreign

[3] http://dir.yahoo.com/

countries also use more ambiguous TLDs such as .com, .org, and .net. We found a total of 49 countries with P3P-enabled web sites. Most of the non-US P3P policies were hosted in the United Kingdom, where we found a P3P adoption rate of 3.25% across the 3,748 .uk sites in our cache. The top five nations also included Japan, Australia, Canada, and Germany.

We found that European Union websites have P3P policies describing more privacy protective practices than US web sites and non-EU websites outside the US, perhaps due to Directive 95/46/EC of the European Parliament which specifies protections for personal data privacy. We found that E.U. websites collect significantly less data from visitors. Websites hosted in E.U. countries are likely to collect half as many types of data as non-E.U. countries. To a lesser extent, E.U. websites are also less likely to use data for marketing, tailoring, and research. However, most significantly we found that E.U. websites are less likely to share data as well as more likely to provide consumers with access to the data collected on them.

A similar study used a geographical location service to determine the nationality of P3P sites rather than TLDs [29]. They also found P3P adoption in 49 nations, although these are not necessarily the same nations. Their method differed from ours in that they used Alexa Language lists to identify sources of non-US P3P policies, whereas we used the Privacy Finder cache. The sites in the Privacy Finder cache were compiled as a result of mostly English-language searches. Thus, our method risks under-counting non-US policies and over counting policies in English speaking countries.

The study based on the Alexa lists found that P3P adoption in the United Kingdom is about 3 times the rate as in the United States (34.4% v. 11.4% in February 2005) [29]. That finding is considerably different than what we found. We suspect the differences may be due to biases in the Alexa lists or the fact that the Alexa lists included only 32 UK web sites. Further work is needed to develop a P3P survey methodology that samples websites more consistently across countries.

5 Longitudinal Trends

In Section 4 we present snapshots of P3P deployment at the time we collected our data. However, P3P deployment is not static. P3P policies are added, removed, and changed on a regular basis. In this section we track P3P deployment rates over a two-and-a-half-year period among over 5,000 websites on 10 lists of URLs. We also examine P3P deployment rates over a one-year period for 30,000 popular domains. Finally, we examine P3P policy additions, deletions and changes associated with websites in the Privacy Finder cache over an eight-week period.

5.1 P3P Deployment Trends for Selected Site Lists: Summer 2003 to Winter 2006

In the summer of 2003, Byers, et al. conducted the first automated study of P3P adoption [30]. 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 website privacy policies. These lists consisted of popular websites, a random sampling of websites, and a refined list of random websites [31]. 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 [32,33]. 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. They used Froogle to create a list of 1,017 commerce-related sites [24]. They used Yahooligans!, a web index run by Yahoo! and geared towards children of ages 7-12, to create a list containing 900 sites. They crawled FirstGov to create a list of 344 U.S. government websites. Finally, they crawled Google News 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 6.

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 2006. The results here show that overall there was an increase in total P3P adoption over the two-and-a-half year period. The total percentage of sites with P3P policies 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 websites. This increase is probably due to the E-Government Act of 2002 which mandates government agencies post machine-readable privacy policies on their websites [34]. Additional increases can be seen with regard to news-related sites as well as websites targeted at children.

A recent study of P3P adoption from the University of Alberta had some overlap with our study. The Alberta study had similar findings on longitudinal adoption rates in the areas of overlap. For example the Alberta researchers

	Number in list	Sites reached in 2003	P3P-enabled in 2003	Sites reached in 2006	P3P-enabled in 2006	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 6
Revisiting the 2003 study on P3P-adoption. This study examined lists of websites from the Progress and Freedom Foundation, comScore, Alexa, FirstGov, Google, and Yahoo!

found a statistically significant increase in government adoption of P3P, which is in keeping with our findings. They also found statistical significance for an increase in P3P adoption from sites with BBBOnline privacy seals, which we did not study [29]. Taken together, these two studies demonstrate that P3P adoption rates vary considerably across industries. Estimates of P3P adoption are also very sensitive to the methodology used to select web sites for examination.

5.2 P3P Deployment Trends for 30,000 Most Popular Domains: December 2005–December 2006

Our longitudinal study found an overall increase in P3P adoption across the 10 site lists over the two-and-a-half year period, but it also showed some areas where adoption is decreasing. It is not clear which lists give the clearest picture of P3P adoption trends. As a benchmark for this study, we examined the 5,856 URLs used in the 2003 study [30], against our database of search results to develop an understanding of how often high traffic websites appear in search results. Of our 1,122,643 hits, we found that 331,943 (29.57%) correspond to the 5,856 websites in this list. This indicates that when users use search engines, they are presented with sites that are not on this list over seventy percent of the time. Likewise, research has shown that most search engine results do not appear on lists of popular websites [35]. Therefore, examining search engine results may yield data that is more applicable to the user experience than using lists of the most popular websites.

To understand P3P adoption trends among search engine results, we checked the list of the 30,000 most clicked on domains from AOL search results for P3P policies in December 2005 and again in December 2006. We found that 2,564 domains (8.54%) had P3P policies in 2005 and 2,934 domains (9.78%) had P3P policies in 2006. This represents a 14.43% increase in P3P adoption over a one-year period among the 30,000 sites that users most often click on when searching the web.

5.3 P3P Policy Additions, Deletions, and Changes

We used our Privacy Finder cache to monitor P3P policy additions, deletions, and changes over an eight-week period beginning October 25th, 2006 and ending on December 20th, 2006. We examined approximately 9,000 P3P-enabled websites on a daily basis to track the rate of changes made to P3P policies. We also examined approximately 175,000 other websites without P3P on a weekly basis to determine if they added new P3P policies. Note: these numbers are approximate because the size of our cache increased throughout the study period.

5.3.1 Policies Added

During the study period we observed 470 new policies added to the approximately 175,000 websites we monitored, an average of 59 per week. As companies often own multiple websites that have the same privacy practices, the same P3P policy is often used on multiple websites. This set of 470 new policies includes 272 unique policies.

5.3.2 Policies Removed

During the study period 70 of the P3P policies that had been available at the beginning of the study period were removed or became unavailable for various reasons. In 5 cases the web server on which the policy resided was inaccessible.

We found that 54 of the P3P policies had actually been removed. In addition, 11 of the P3P policies were still on the websites, but could no longer be fetched by a P3P user agent due to the addition of a misconfigured robots.txt file. The robots.txt file is used to limit access to files by web crawlers (e.g., to keep a file out of Google's search database). However, if the P3P policy is in a restricted directory, then user agents can no longer access the policy. It seems unlikely people are intentionally going to the effort to create P3P policies and then making them inaccessible. It is more likely they do not understand they need to white list their P3P policy in their robots.txt file. We also discovered an additional 46 policies that appeared to have been removed but were actually still accessible when we checked them later. This indicates that these sites were temporarily inaccessible when Privacy Finder checked them.

As a result of the P3P policies added and removed (including those that became permanently unavailable), the total number of P3P policies available increased by 400 during our study period, an average net increase of 50 policies per week. This reflects a net growth rate of roughly 4.16%. Extrapolating over a year, we predict an increase in P3P deployment of 27% for the websites in the Privacy Finder cache. This is about twice the growth rate we observed during the prior year for the 30,000 most clicked on domains in AOL search results.

5.3.3 Policies Changed

During the study period we saw sixty-nine changes to P3P policies. These changes occurred on thirty-eight different policies. This establishes that at least some P3P policies are not "write once" documents, but rather documents that are updated as conditions change. The changes that we observed fell into four categories: genuine policy changes, contact information changes, syntax changes, and wording changes.

The genuine policy changes are most important, as they impact the privacy practices that users will encounter. We observed policy changes on eight sites. For example, one website switched from BBBOnline to TRUSTe for resolving disputes, and other websites stopped requiring certain types of information to complete a transaction. All of the policy changes we observed improved the overall level of privacy protections offered by those sites. However, we would need to observe changes over a longer period to see whether this is a general trend.

We observed at least thirty sites that provided updated contact information. This information ranged from new email addresses for customer service to different URLs for opting-out or for the natural language version of their privacy policies.

Three of the policies were previously not compliant with the P3P specification; we observed syntax changes made to these policies which made them compliant. These amounted to updating the namespace to match the current version of the schema, as well as adding required XML tags that were previously missing.

We observed wording changes to a dozen P3P policies. These changes took place in the optional natural language descriptions of various elements and do not have an impact on the semantics of the policies. A few of the sites made changes like this multiple times.

It would be interesting to determine if P3P policies are updated more, less, or as frequently as human-readable privacy policies. This is an area for further study.

6 Policy Content

The main purpose of P3P user agents is to provide users with information about the privacy practices of the websites they visit so that they can choose websites that match their preferences. A P3P-enabled search engine can make it easy for users to find the sites with the best privacy policies that have the information or products they want. However, P3P-enabled search engines are not all that useful if few or none of the sites returned have P3P policies available, or if users find the privacy practices of the P3P-enabled sites returned to be unacceptable. In Section 4 we discussed the number of P3P-enabled search results that users can expect. In this section we discuss the types of privacy practices users are likely to find in those results.

6.1 Privacy Finder Settings, Marketing, and Sharing

We used a set of five APPEL files discussed in Section 3 to examine the compliance of P3P-enabled websites with the three Privacy Finder settings and to determine whether they engage in marketing practices using personal information (excluding opt-in marketing) or share personal information with third parties (excluding opt-in sharing, sharing with delivery companies, and sharing with companies acting as agents for the website). Table 7 shows the percentage of P3P-enabled sites for each search engine that resulted in a 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 setting. This is because they either collect

API	Low	Medium	High	Don't Share	Don't 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 the percentage that 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.

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.

Ideally, a Privacy Finder search will yield multiple sites that completely match a user's privacy preference settings (indicated with four green boxes). 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 8. 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 site present in the top ten search results that matched the Privacy Finder low 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.

We observed that Google and AOL were similar in the types of policies that they returned, while 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 8, Google is also more likely to return a larger number of hits that match a user's preference settings.

Table 9 compares the types of policies found across the Yahoo and Google search APIs using the Froogle data. This is similar to the data depicted in Table 7; in almost all cases Google yields "better" policies, since the websites returned are less likely to share or analyze personal information as well as engage in marketing practices. When comparing Tables 7 and 9 we see that typical searches are more likely than e-commerce searches to return sites that

Google

Hits	Low	Medium	High	Don't Share	Don't 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	Don't Share	Don't 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	Don't Share	Don't 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 8
This table shows the cumulative frequency of P3P-enabled search hits that 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.

API	Low	Medium	High	Don't Share	Don't Market
Google	64.23%	54.98%	22.83%	67.79%	55.77%
Yahoo!	59.16%	48.30%	29.39%	59.34%	47.43%

Table 9
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 the percentage that complied with each preference level. In all cases the differences between the two search engines are significant ($p < 0.0005$).

share data with other companies. In addition, typical searches are less likely to return sites that engage in marketing. One possible explanation is that the sites returned using the Froogle data are more likely to be commerce websites that collect data to complete a purchase, whereas the AOL 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.

6.2 Industry Trends

We examined the content of the P3P policies for 1,181 websites found in the nine Yahoo! Directory categories in detail, focusing on types of data collected, data use, and data recipients.

6.2.1 Types of Data Collected

Figures 4 and 5 show that most websites in almost every category collect computer information, demographic information, interactive data, navigation information, online contact information, physical contact information, and unique identifiers. Shopping sites tend to collect data in most categories, while government sites tend to collect data in few categories.

Perhaps most surprising is how many shopping sites claim to collect political information. Also, nearly a quarter of banking websites report they collect location data. Location data refers to the real-time location of people using the website, perhaps based on a GPS reading. It is possible that banks misunderstood, and took home address to be location data. Or, if accurate, this may be a reflection of the increased use of geoIP databases to determine the physical location of an IP address. Such information could be useful for fraud detection, but we have no confirmation that this is really happening. It is possible that the use of these P3P tags, `<political/>` and `<location/>`,

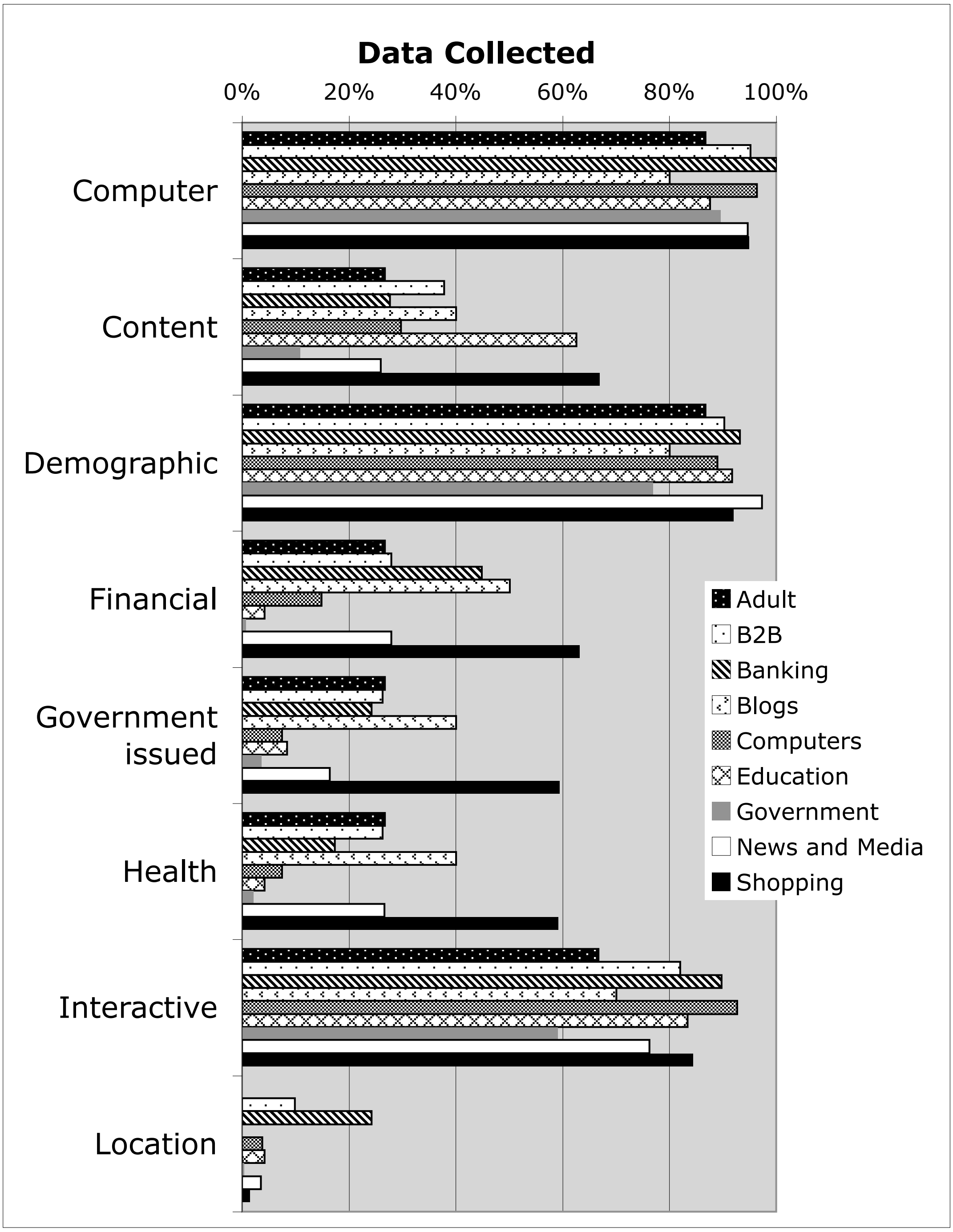

Fig. 4. Data collection by Yahoo! Directory categories. Part 1 of 2.

are in error. Likewise it is likely that all sites actually collect navigation and computer data, but some of them incorrectly fail to disclose this in their P3P policies, probably due to a misunderstanding of the P3P specification. This phenomenon is further explored in Section 7.2.1.2.

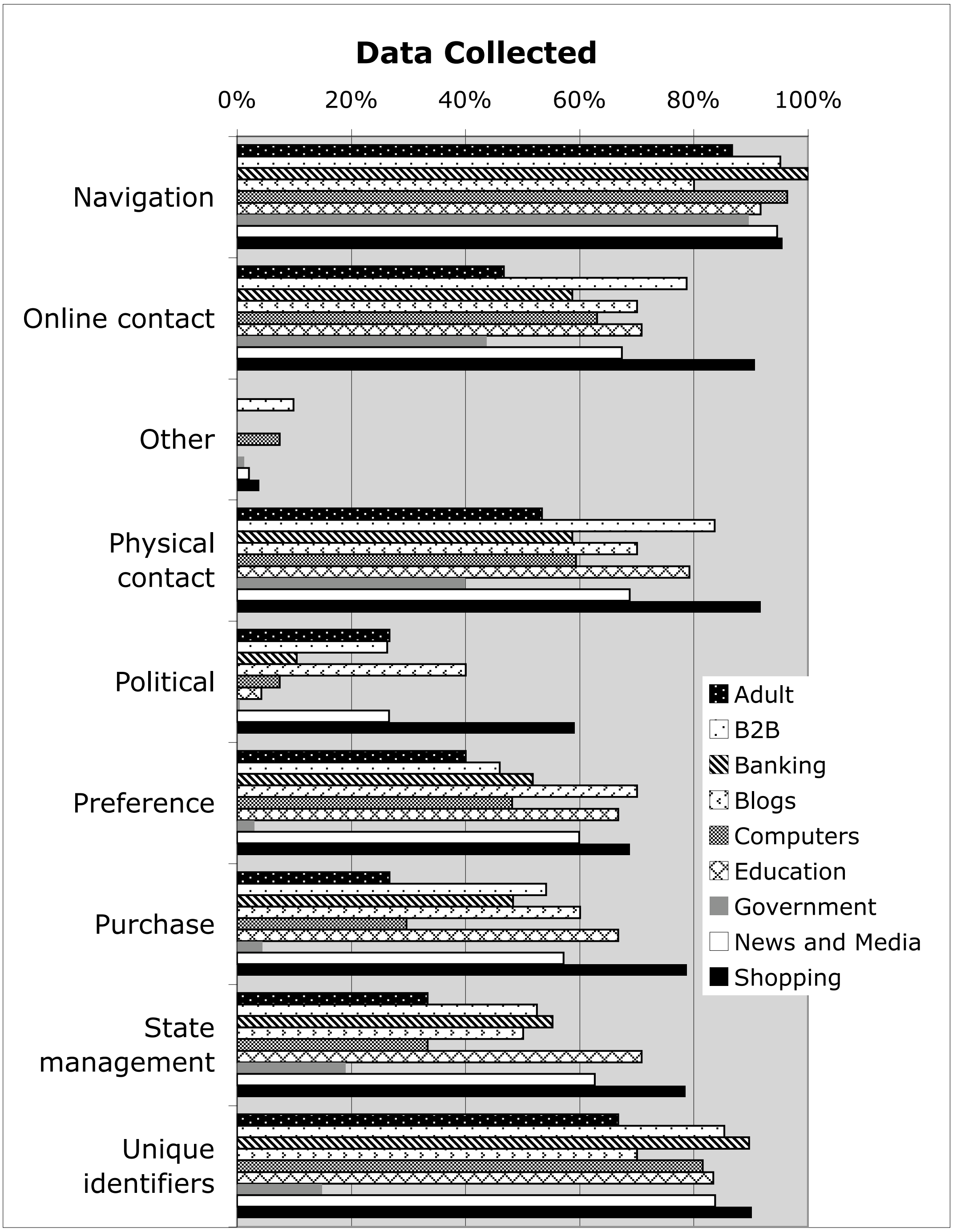

Fig. 5. Data collection by Yahoo! Directory categories. Part 2 of 2.

6.2.2 Uses for Data Collected

Figures 6 and 7 show how collected data is being used across each P3P category. Support of the current activity, system administration, and research and development are the three most prevalent uses for data. Shopping sites use data in just about every way possible, whereas government sites are more restrictive. Use of information for marketing varies considerably across categories, with telemarketing less prevalent than other forms of marketing. Many sites profile individuals, either for their own analysis purposes or for making decisions that impact the individuals. More sites use pseudonymous profiles (`<pseudo-analysis/>` and `<pseudo-decisions/>`) than identified profiles (`<individual-analysis/>` and `<individual-decisions/>`).

We were surprised to see non-government sites report that they store data for "historical preservation." It is likely that many of these sites misunderstood the proper usage of the `<historical/>` tag.[4]

6.2.3 Data Recipients

Figure 8 shows with whom data is shared by websites in each category. Shopping sites share—or sell—data more widely than any other sector, while government sites do not share data very frequently.

6.3 Popular Sites

We examined the 21 P3P-enabled sites on the Popular list (described in Section 3) and compared these sites with 7,741 P3P-enabled sites in the Privacy Finder cache[5] in order to determine whether popular sites have significantly different privacy practices than other sites.

We used our set of 67 APPEL files to compare the two sets of P3P policies with regard to the types of information that may be collected (`<CATEGORIES>`), how information may be used (`<PURPOSE>`), how information may be shared (`<RECIPIENT>`), information about an individual's ability to access his or her own information in the site's records (`<ACCESS>`), data retention policies (`<RETENTION>`), and options for dispute resolution (`<DISPUTES>`).

We determined how many sites in each set engaged in each data practice and used a chi-square test to examine the differences between the two data sets. We found no significant differences between the P3P policies in the two data sets in any of the areas examined. We can conclude that there is not a

[4] P3P1.0 defines `<historical>` as follows: "Historical Preservation: Information may be archived or stored for the purpose of preserving social history as governed by an existing law or policy. This law or policy MUST be referenced in the `<DISPUTES>` element and MUST include a specific definition of the type of qualified researcher who can access the information, where this information will be stored and specifically how this collection advances the preservation of history" [7].

[5] We used all P3P-enabled sites in the cache at the time of the analysis except those with critical errors.

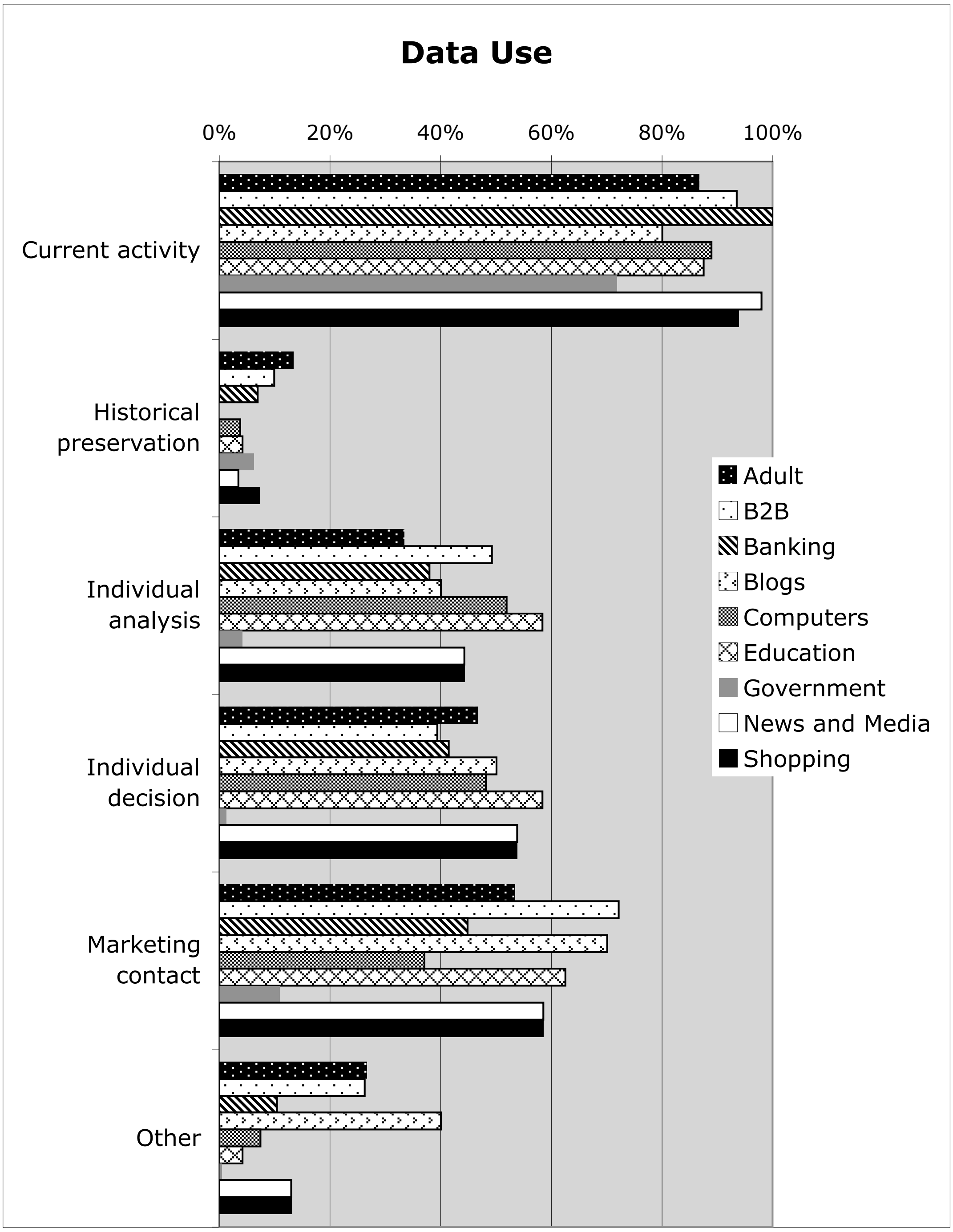

Fig. 6. Data use by Yahoo! Directory categories. Part 1 of 2.

significant difference in policies between sites on the Popular list and other P3P-enabled sites. This suggests that while P3P deployment rates vary significantly according to website popularity, the content of P3P policies does not.

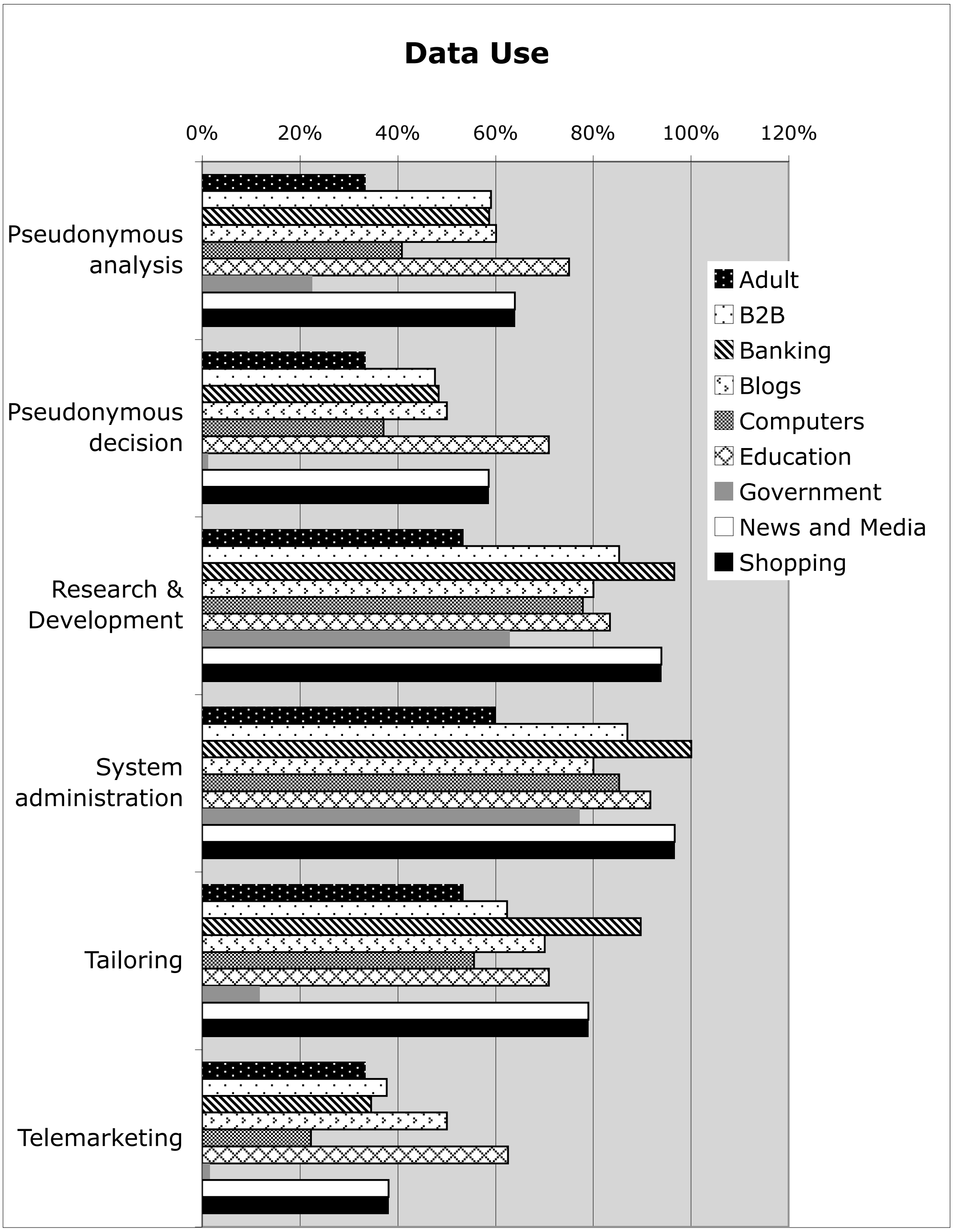

Fig. 7. Data use by Yahoo! Directory categories. Part 2 of 2.

7 Policy Errors

There are two types of errors people can make while coding P3P policies: semantic errors and syntactic errors. Semantic errors occur when the P3P policy complies with the P3P specification, but does not accurately reflect the site's natural language policy. For instance, they may mistakenly claim they

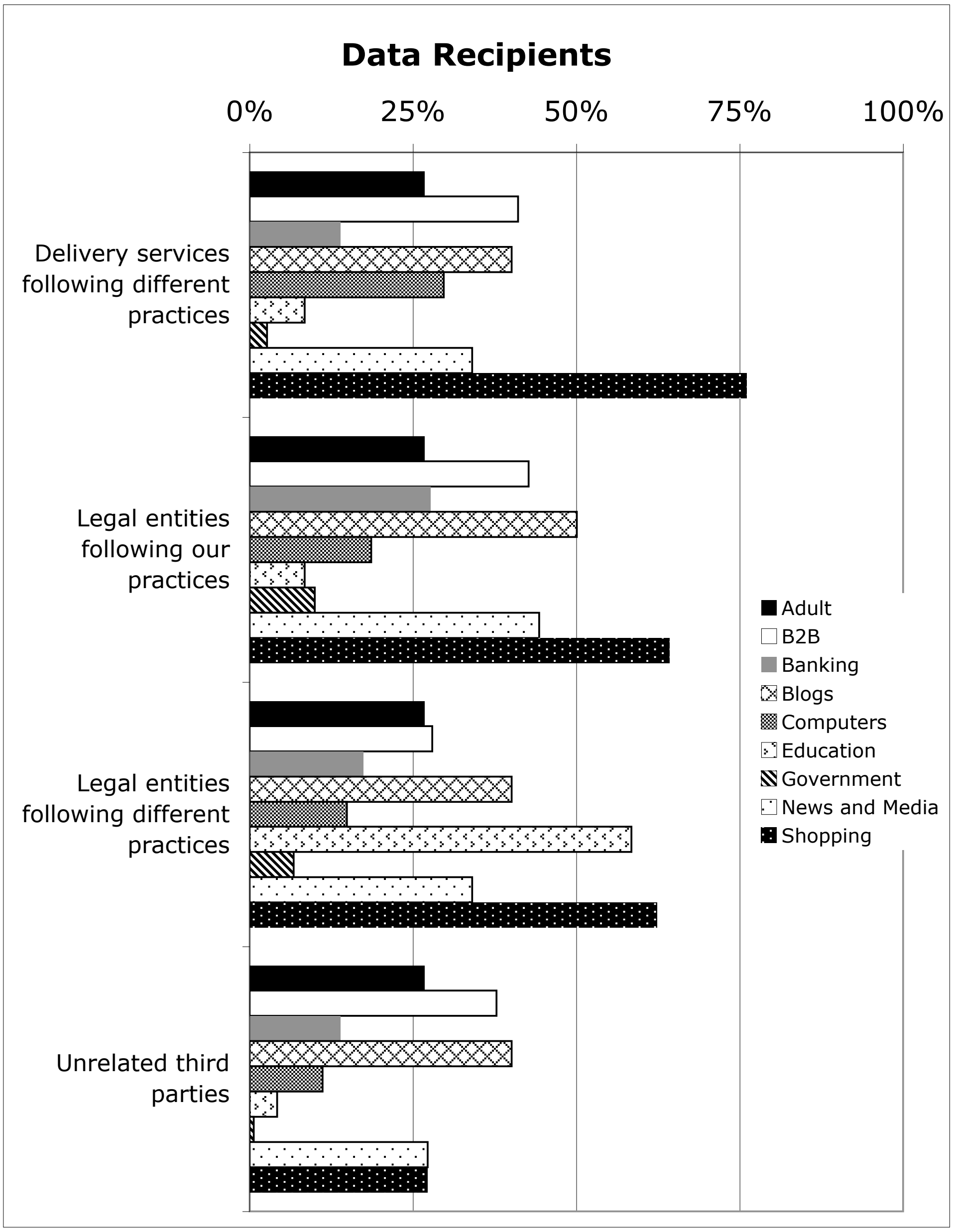

Fig. 8. Data recipients by Yahoo! Directory categories.

retain data for 30 days when they have off-site backups for a year. We are able to detect semantic errors by comparing P3P policies to human-readable policies. If they do not agree, clearly something is wrong, though we often cannot tell which policy is accurate. (We also cannot detect errors that occur in both the human-readable and P3P policies.) Syntax errors occur when the P3P policy that is published does not comply with the P3P specification. In some of these cases this makes it impossible to parse the policy, while in other cases the policy can still be parsed.

Critical syntax errors can prevent a policy from being evaluated and thus render it invalid. Semantic errors can create liability problems for a website. The U.S. Federal Trade Commission (FTC) is charged with protecting against "unfair and deceptive practices." [6] A P3P policy that states something other than what is stated in the natural language privacy policy may be interpreted as a deceptive business practice. Thus, a P3P policy with semantic errors may subject a U.S. website to FTC enforcement.

7.1 Syntactic Errors

When we attempted to validate the P3P policies we had collected, we found that the majority of these policies contained syntax errors. In 2003, 33% of the sites discovered contained errors as found by the W3C P3P Validator [30]. However, when using the W3C validator in 2006, we discovered that only 27% of the total sites examined did not contain any errors. It is possible that changes made to the validator since 2003 have resulted in the detection of errors that previously existed but were not detected. Another study of syntax errors in P3P policies conducted in November of 2005 found error rates as small as 13% among a list of U.S. government websites, to as much as 69% among a list of P3P-enabled sites displaying the BBBOnline web seal [29]. While our results are not directly comparable because they segmented websites into categories while we looked at aggregate rates, our overall error rate is higher than any they report.

Most of the errors in this study were considered "non-critical errors" in that they conflicted with the P3P specification, but the evaluator was still able to function correctly. These errors usually amounted to using an older version of P3P. This type of error can be corrected easily. 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). About nine percent of the P3P policies we evaluated in 2006 had critical errors, while six percent of the policies evaluated in the 2003 study had critical errors.

7.1.1 Types of Syntactic Errors

We used Perl code from the W3C's P3P validator as the basis for our own automated validator [14]. Using our validator, we were able to classify P3P syntax errors into the following fourteen categories:

(1) **Old Version** – P3P policy or policy reference file are based on a pre-release version of the P3P specification rather than the final P3P 1.0

[6] 15 U.S.C. §45(a)

Recommendation.

(2) **No Policy Name** – P3P policy and/or policy reference file do not include proper policy names. While technically an error, this usually only causes problems for some websites with multiple P3P policies. This problem usually occurs when policies are based on a pre-release version of the P3P specification.

(3) **Policy Validation Error** – P3P policy or policy reference file is missing required elements or has other errors that prevent it from being validated. This error usually prevents policy evaluation.

(4) **Bad XML Root** – P3P policy or policy reference file has an invalid XML root node, which prevents the file from being parsed. This error prevents policy evaluation.

(5) **Policy Expired** – P3P policy or policy reference file has an explicit expiration date that is in the past.

(6) **Policy Vocabulary Error** – P3P policy contains unrecognized elements or improperly references data elements. This error usually prevents policy evaluation.

(7) **No Policy Elements** – P3P policy file is blank, does not contain a policy of the name specified in the policy reference file, or cannot be parsed into XML. This error prevents policy evaluation.

(8) **Incorrect XML** – P3P policy or policy reference file are not valid XML documents. This error prevents policy evaluation.

(9) **Policy Access Error** – P3P policy file cannot be accessed (HTTP 404, 403, etc.). This error prevents policy evaluation.

(10) **No Namespace** – P3P policy or policy reference file does not include the P3P version number. This error prevents policy evaluation.

(11) **Malformed INCLUDE/EXCLUDE** – Policy reference file has invalid or missing INCLUDE elements. This makes it impossible to determine what parts of the website the P3P policy covers. This error prevents policy evaluation.

(12) **No `<META/>` Tag** – Policy reference file does not begin with required `<META/>` tag. This error usually prevents policy evaluation.

(13) **No Policy Found** – A policy reference file exists but the P3P policy it references does not exist. This error prevents policy evaluation.

(14) **Not A Policy** – P3P policy file can be located and appears to be XML, but is unrecognizable. This is usually because unknown tags have been included before the `<POLICIES>` or `<POLICY>` tags. This error prevents policy evaluation.

Some of these errors are considered critical errors—errors that prevent the policy from being evaluated, while others are considered non-critical errors. We observed that syntactic errors were more prevalent across less popular sites and that the more popular sites (selected from the Popular list) were less likely than other P3P-enabled sites to contain critical errors.

7.1.2 Errors in Popular Sites

We first examined the P3P policies of the 21 P3P-enabled sites on the Popular sites list. There were only six policies (28.6%) that did not contain any errors. However, most of the errors that did exist were trivial and only one site, qvc.com, had a P3P policy that had critical errors. The error in qvc.com's policy was that the policy reference file referred to a policy URL that did not exist (an HTTP 404 error occurred when we attempted to retrieve this policy).

The most prevalent error was the use of an old namespace. All P3P policies should be using the current XML namespace, *http://www.w3.org/2002/01/P3Pv1*. The previous namespace, *http://www.w3.org/2001/09/P3Pv1*, was created before the P3P 1.0 specification was finalized in 2002, but is still used by many websites. It is possible that some of these websites were early adopters of P3P who have yet to update their P3P policies. Fifteen of the twenty-one (71.4%) sites that we examined were using an old namespace. Fortunately this error is non-critical, and while a departure from the specification, usually does not prevent a P3P evaluator from parsing a policy.

The next most prevalent error was the use of an incorrect XML root element. All P3P policy files must start with either `<POLICIES/>` (for a stand-alone policy) or `<META/>` (for a policy embedded in a policy reference file). This type of error could potentially be critical. However, we have noticed that many policy files incorrectly use the `<POLICY/>` tag at the beginning (as had been specified in an early draft of the P3P specification), and have thus adapted our validator to recover from this type of error. Eight policies in the set contained this error.

The other non-critical syntax errors that we found all relate to the name of the policy. According to the P3P specification, every policy must have a name. This is so that when multiple policies are used, the parser can automatically locate the most appropriate policy. However, if a site has only one policy, this error does not pose a problem. Among the Popular sites we found that three policies did not include a policy name, and that one policy included an invalid name. The policy with an invalid name, usps.com, contained multiple spaces, which are not permitted by the P3P schema.

7.1.3 Errors in Other Sites

We compared the syntactic error rates of the Popular sites with the error rates of P3P-enabled sites in our Privacy Finder cache. At the time of this analysis the cache contained 14,720 P3P policies, of which 10,706 (73%) contained errors and 1,306 (9%) contained critical errors. Table 10 shows a summary of these errors, as well as a comparison with the error rates found among the

Error	**Top 100**	**Privacy Finder**
Old Version	15 (71.4%)	9,155 (62.2%)
No Policy Name	3 (14.3%)	6,289 (42.7%)
No Errors	6 (28.6%)	4,014 (27.3%)
Policy Validation Error	1 (4.8%)	1,157 (7.9%)
Bad XML Root	8 (38.1%)	1,125 (7.6%)
Policy Expired	0	474 (3.2%)
Policy Vocabulary Error	0	453 (3.1%)
No Policy Elements	0	252 (1.7%)
Incorrect XML	0	204 (1.4%)
Policy Access Error	1 (4.8%)	183 (1.2%)
No Namespace	0	151 (1.0%)
Malformed INCLUDE/EXCLUDE	0	56 (0.4%)
No `<META/>` Tag	0	21 (0.1%)
No Policy Found	0	5 (0%)
Not A Policy	0	2 (0%)
Total Policies	21	14,720

Table 10
Comparison of the syntax errors found in the P3P policies of the Popular sites with the policies found in the Privacy Finder cache.

Popular sites.

It can be seen that the P3P policies suffer from the same common errors. Unfortunately, the sample size of P3P-enabled sites from the Popular List was so small that we were only able to make two significant comparisons. Upon performing a z-test for proportions, we found that the sites stored in Privacy Finder's cache were significantly more likely to be missing policy names ($p < 0.018$), while the popular sites were significantly more likely to have an incorrect XML root element ($p < 0.0005$). All other differences between the two lists were insignificant.

7.2 Semantic Errors

In addition to errors in P3P syntax, we also examined P3P policies for semantic errors. We considered conflicts between P3P policies and their corresponding natural language privacy policies to be semantic errors in the P3P policies. However, it is possible that in some cases the P3P policies are correct and the errors are actually in the natural language policies.

7.2.1 Types of Semantic Errors

We used the sixty-seven APPEL files discussed in Section 3 to evaluate each of the 21 P3P policies from the Popular list. We then used these files to evaluate the pseudo-P3P policies that had been created for each site by our coders based on the corresponding natural language policies. By comparing the results of the APPEL evaluations for each P3P policy with the results of the evaluations for the corresponding natural language policy, we were able to find semantic errors. Table 11 shows the twenty-one sites whose policies we examined, as well as the number of errors each one contained.

As can be seen from the table, we encountered multiple conflicts with every policy that we examined. However, some policies had far more errors than others. There are a number of reasons for policy disagreements. In some cases, the P3P policies are clearly incorrect. In other cases, it is possible that the natural language policies are overly vague. And it is also possible that some of these conflicts stem from perceived ambiguities in the P3P specification. For instance, `<interactive/>`, `<navigation/>`, and `<computer/>` can all apply to data that is transmitted within HTTP headers.

Table 11 shows that some of the policy areas are easier to make mistakes in than others. The table shows the number of possible mistakes in each area, which is based on the number of possible P3P elements. However, the number of mistakes actually made are not evenly distributed across the elements. For instance, a higher proportion of mistakes were made with regard to why data is collected (the `<PURPOSE>` tag), than with the types of data collected (the `<CATEGORIES>` tag).

7.2.1.1 `<ACCESS>` Errors. Errors using the `<ACCESS>` element were found in nine of the twenty-one policies. The P3P specification specifies six possible mutually-exclusive `<ACCESS>` tags: one for sites that do not collect personal information, one for sites that do not provide any access, and four for sites that provide access to some or all of a user's personal information. The human-readable policy and P3P policy might both state that access is provided, but may disagree on the extent of access (for example, the natural language policy might say that all information can be accessed, while the P3P policy might state that only contact information can be accessed).

	<ACCESS> (1)	<CATEGORIES> (17)	<DISPUTES> (4)	<NON-IDENTIFIABLE> (1)	<PURPOSE> (12)	<RECIPIENT> (6)	<REMEDIES> (1)	<RETENTION> (1)	Total
1. yahoo.com	0	3	0	0	2	4	0	0	9
2. geocities.com	0	3	0	0	2	4	0	0	9
3. hotmail.com	1	1	0	0	2	0	0	1	5
4. superpages.com	1	6	0	0	5	3	0	1	16
5. angelfire.com	0	0	0	0	3	1	0	0	4
6. walmart.com	0	4	1	0	4	2	1	1	13
7. go.com	0	1	0	0	3	2	0	1	7
8. microsoft.com	0	2	0	0	2	0	0	0	4
9. ticketmaster.com	1	7	0	1	5	3	0	0	17
10. usps.com	0	1	0	0	3	2	1	1	8
11. dealtime.com	1	7	1	0	5	1	1	1	17
12. rootsweb.com	1	5	0	0	5	2	0	1	14
13. hgtv.com	1	3	0	0	1	3	0	1	9
14. wachovia.com	0	5	0	0	5	1	1	1	13
15. tripod.com	0	0	0	0	3	1	0	0	4
16. sportsline.com	0	6	0	0	2	3	0	1	12
17. qvc.com	1	7	0	0	4	1	0	0	13
18. download.com [7]	0	5	1	0	2	5	0	0	13
19. usatoday.com	0	2	1	0	1	1	1	0	6
20. about.com	1	4	2	0	4	2	0	1	14
21. wunderground.com	1	7	0	0	3	0	0	0	11
Policies with Error	**9**	**19**	**5**	**1**	**21**	**18**	**5**	**11**	**217**

Table 11
Semantic error rates among the 21 most popular P3P-enabled websites. The top row shows the major policy elements, with the number in parentheses denoting the number of possible errors associated with that element (e.g. there are seventeen different <CATEGORIES> elements, and policies may use any combination; whereas <ACCESS> has six mutually-exclusive elements). It should be noted that geocities.com and yahoo.com use the same policy, as do angelfire.com and tripod.com.

7.2.1.2 `<CATEGORIES>` Errors. Only two P3P policies correctly specified the types of data that were being collected. Eighty percent of the `<CATEGORIES>` errors were due to sites omitting data types from their P3P policy that were mentioned in their natural language policies. Thus, users reading only a P3P policy might be surprised to find a site collecting more data than what was advertised. Many of these errors may stem from a misunderstanding of P3P categories. For instance, the `<content/>` category is used when a site collects user-generated content, such as posts to forums or message boards. Ten sites mention the collection of such content in their natural language policies, yet fail to mention it in their P3P policies. In some cases, we observed errors that were unlikely to have stemmed from a misunderstanding of the P3P specification. For instance, wachovia.com, a bank that allows individuals to open accounts online, has a P3P policy that claims they do not collect government issued-identification (e.g. Social Security Numbers) or any contact information.

7.2.1.3 `<DISPUTES>` Errors. Only five policies had `<DISPUTES>` errors. Four of the P3P policies failed to provide customer service contact information that was provided in a natural language policy. Two sites mentioned an independent organization (e.g. TRUSTe) in one of the policies, but not the other. These errors are unlikely to mislead users about a website's privacy practices.

7.2.1.4 `<NON-IDENTIFIABLE>` Errors The `<NON-IDENTIFIABLE>` element is used to indicate that no personally identifiable information is collected by the website. Using this element allows the policy writer to omit certain other tags, since if no information is collected, a description of how information is used is unnecessary. However, very few sites can legitimately use this tag, since most of them log IP addresses, which are considered to be potentially identifiable information. Only one site, ticketmaster.com, used this element in their P3P policy. This is a clear error as users can purchase tickets through the website and are thus required to enter contact and billing information.

7.2.1.5 `<PURPOSE>` Errors. The `<PURPOSE>` element specifies the ways in which collected data may be used. We found more discrepancies between the natural language and P3P policies for this element than for any other element. Such errors were made in all 21 of the policies we examined.

In some cases `<PURPOSE>` errors can be quite misleading. For example eight natural language policies (about.com, dealtime.com, qvc.com, rootsweb.com, sportsline.com, superpages.com, ticketmaster.com, and wachovia.com) mention that they may contact individuals for marketing by means other than telephone, while their corresponding P3P policies do not mention any contact. We also observed the opposite problem, where marketing contact is reported in P3P policies, but not in the corresponding natural language policies. None of the natural language policies that we examined make any mention of telemarketing, yet five P3P policies claim to engage in telemarketing on either an opt-out basis (hotmail.com and microsoft.com), or require it without any consent (geocities.com, wunderground.com, and yahoo.com). In one case (wunderground.com), the P3P policy states that individuals may be contacted via a means other than telephone; however, the corresponding natural language policy makes no mention of this. It is hard to explain away these sorts of policy differences as a misunderstanding of P3P, as the descriptions of the `<contact/>` and `<telemarketing/>` elements are rather straightforward.

The most common `<PURPOSE>` error we observed was incorrect use of the customization and analysis purposes, which are recognized to be confusing.[8] The P3P specification distinguishes between customization that involves creating a user profile and customization that does not involve creating a user profile, between identified and pseudonymous profiles, and between profiling for analysis purposes and profiling to make decisions that will impact the user. Forty-seven discrepancies—over seventy percent of the `<PURPOSE>` errors—involve the use of these elements. Thirty-three of these errors involve omitting some of these purposes in the P3P policies, while the other fourteen are due to reporting practices in the P3P policies that are not mentioned in the natural language policies.

7.2.1.6 `<RECIPIENT>` **Errors.** The differences between the P3P and natural language policies with regard to data recipients were the most significant of any element ($\chi^2 = 17.32$, $df = 4$, $p < 0.01$). This is particularly troubling as web users generally read privacy policies in an attempt to determine data sharing policies [16]. Overall, 41 errors were made across the six elements in this category. In 28 cases (68%), the natural language policy states that data may be shared with recipients who are not specified in the corresponding P3P policy. Only six of the websites examined either accurately report their data sharing policies (hotmail.com, microsoft.com, and wunderground.com) or their P3P policies are overly inclusive (geocities.com, usatoday.com, and yahoo.com) in their reporting of data sharing.

Eleven websites (about.com, angelfire.com, dealtime.com, qvc.com, rootsweb.com, sportsline.com, superpages.com, ticketmaster.com, usps.com, wachovia.com, and walmart.com) stated that they share data with third parties in their natural language policies but do not mention this in their P3P policies, although

[8] Cranor provides advice on distinguishing these purposes on p. 94-95 of *Web Privacy with P3P* [36].

in some cases the data sharing mentioned in the natural language policy is by opt-in only. In most cases it is hard to attribute this error to a misunderstanding of the P3P specification.

Many websites fail to use the `<public/>` element to disclose that data may be posted on public forums. Nine sites mentioned public forums in their natural language policies, yet failed to disclose them in their P3P policies.

Errors involving the `<delivery/>` element may be due to confusion about how this element should be used, or perhaps confusion about the privacy practices of delivery companies. The `<delivery/>` element indicates that data may be shared with delivery services, and that the delivery services may use this data for additional purposes. In the corresponding natural language policies, four sites claim that data is only shared with delivery services in order to complete a transaction, and that the data is not used for any other purposes. If this is the true policy, the `<delivery/>` element need not be used. However, some popular American delivery companies do not commit to using delivery address data only for delivery purposes. For example the UPS privacy policy states, "We use information about our customers, their packages, and their shipping activity to provide or enhance the services we make available to our customers, communicate with our customers about additional services they may find of value...." Thus, it may be the natural language policy that is in error.

7.2.1.7 `<RETENTION>` Errors. With the exception of a few industry-specific regulations, there exist few legally-binding guidelines as to what elements must be included within an American website's privacy policy. However, to comply with the P3P specification, certain practices must be disclosed. One specific example is data retention. To comply with the specification, a P3P policy must specify the length of time that personally identified information is retained. It appears that P3P has prompted many companies to disclose their data retention policies when they otherwise might not do so. Of the twenty-one sites examined, twelve sites did not mention their retention policy within their natural language policies. However, these twelve sites did mention data retention in their P3P policies (as required). In this example we can see that P3P is serving users by forcing companies to disclose information they might otherwise not disclose.

We did encounter some `<RETENTION>` errors. According to the P3P specification, if the natural language policy does not specify a data destruction timetable, then data is assumed to be stored indefinitely. If any other data retention elements are used besides `<indefinitely/>` or `<no-retention/>`, then the corresponding natural language privacy policy must specify a data destruction timetable. We discovered that none of the natural language policies we examined outlined a specific data destruction timetable, yet eleven sites used tags that require such a timetable.

7.3 Policy Examples

From our examination of individual P3P policies, we observed that some policies seem to suffer mostly from a few minor errors in interpretation of the P3P specification (e.g. the policy at hotmail.com). Other P3P policies have discrepancies between P3P and natural language policies that are likely intended to ensure that the P3P policy is broadly inclusive of many sites' privacy practices (e.g. the policy at yahoo.com). Occasionally, P3P policies contain many significant errors and may result from a total misunderstanding of P3P (e.g. the policy at wachovia.com).

Hotmail.com's P3P policy states that access is given to "contact and other information." However, the natural language policy claims that access will be given to *all* personally identified information. The P3P specification provides a tag to specify all personal information, yet the authors of this particular policy chose not to use it, a possible oversight. Another example of a possible misunderstanding comes with the use of the `<financial/>` tag, which is used for collecting information beyond what is needed to facilitate a purchase—such as account balances, financial history, etc. The natural language policy only made mention of purchase information, but this tag was present in the P3P policy. While these inconsistencies raised errors during our analysis, they did not change the overall "level" of privacy afforded by either policy.

On the other hand, we found that the yahoo.com P3P policy covers far more than their natural language policy. The P3P policy claims that health information and political information may be collected by yahoo.com, however the natural language policy makes no mention of this. We also saw this same phenomenon with regard to data recipients. Yahoo!'s P3P policy states that data could be shared with delivery services for purposes other than shipment of merchandise, with affiliates for unknown reasons, and may be displayed on public forums. None of these are mentioned in the natural language policy. While it is known that Yahoo! does have user-generated content such as message boards, we could not resolve the other discrepancies. We have two theories for this behavior. First, Yahoo! hosts many third party websites and often does data processing, in addition to acting as an intermediary for any data transmitted to these sites. Thus, since Yahoo! might not have a very good idea of the privacy policies of these third parties, the P3P policy is as broad as possible. Another possible explanation is that Yahoo! frequently adds new services to the website, and has therefore created an overly-broad P3P policy so that it does not have to be updated frequently. In any case, Yahoo! is a good example of a P3P policy that is far more inclusive than the natural language policy. Both Microsoft P3P policies (hotmail.com and microsoft.com), wunderground.com, and go.com all exhibited this phenomenon to some extent. While there were discrepancies between the policies, we believe that overall users stand to benefit since they are being given a worst-case scenario of what a company *could* do with their information.

We have seen some benevolent misuses of P3P policies that do not adversely affect end-users. We also encountered examples of gross mistakes that could adversely affect users while creating liability problems for the publisher of the policy. Regulators have stated that P3P policies are just as legally binding as their natural language counterparts.[9] The Financial Modernization Act of 1999, also known as the "Gramm-Leach-Bliley Act," requires institutions in the financial sector to publish privacy policies.[10] Wachovia, a bank, had some serious discrepancies between their P3P policy and their natural language policy. In Section 7.2.1 we discussed some of the discrepancies with the data they claim to collect. We also discovered that their P3P policy claims that they do not contact customers or engage in marketing, while their natural language policy states otherwise. The P3P policy also claims to not use online information to analyze individual user behavior or engage in profiling, while again, the natural language policy claims otherwise. Finally, the P3P policy implies that data will not be shared with any other entities, while the natural language policy claims that data may be shared with affiliates. As a result of these errors, the posted P3P policy appears to comply with the high, medium, and low Privacy Finder settings, whereas a correctly written P3P policy—consistent with the natural language policy—would not be fully compliant with the medium or high settings.

Wachovia's natural language privacy policy[11] includes a section that explains what P3P is and why a company would post a P3P policy. However, we were perplexed to read that "Wachovia does not currently present its privacy policy in the P3P format," despite the fact that they do. An email exchange with Wachovia's customer service department only resulted in their continued denial of currently or ever having a P3P policy. Upon reexamining their website, it appears that this problem was fixed in February 2007. It should also be noted that the link to the natural language policy found within the P3P policy points to a privacy policy that is different from the one linked from the

[9] At the November 2002 W3C Workshop on the Future of P3P, panelists form the European Commission, Ontario Privacy Commissioner, and Office of the New York Attorney General "expressed the opinion that P3P policy statements (in XML) are equally as binding on service operators as are the human-readable policies that websites generally post. Whether a policy is in a machine-readable code that is translated by a user agent, or simply in HTML on a website, the policy constitutes a representation to consumers on which they can be expected to rely" [12].

[10] 15 U.S.C. §6801 et seq.

[11] http://www.wachovia.com/inside/legal_footer/0,,2157_2158,00.html

bottom of every page on the site.

7.4 Privacy Levels

While the number of discrepancies between P3P policies and natural language policies is troubling, many of these errors may have little or no impact on a P3P user agent's behavior. To investigate the extent to which these errors might impact user agent behavior, we evaluated P3P policies and their corresponding natural language policies against the Privacy Finder high, medium, and low settings. Each of these settings represents a composite of multiple elements within a P3P policy, and takes into account only a subset of the elements (those deemed most important to a user selecting that setting).

For the 21 policies we examined, we found six cases where the natural language policies yielded warnings on the highest privacy level, whereas the P3P policy did not, and seven cases where the natural language policy yielded warnings on the medium privacy level, whereas the P3P policy did not. Conversely, there were three cases in which the P3P policy yielded warnings on the highest privacy level but the natural language policies did not. This phenomenon also occurred once each on the medium and low settings. It is not clear whether the P3P policy or natural language policy correctly reflects each website's true policy.

In some cases, companies created overly-inclusive P3P policies. This type of error is fairly harmless as it still allows the user to make an informed decision based on a worst-case scenario. However, many other companies have the opposite problem—creating P3P policies that are far less stringent than their natural language counterparts. This type of error does not serve the user well. However, most of these errors did not impact the overall Privacy Finder privacy level. Thus, despite the high error rate in P3P policies, it still appears to be generally useful when determining whether a site's privacy policy is "good" or "bad."

8 Representativeness of P3P Policies

P3P policies facilitate the rapid analysis of large numbers of privacy policies. However, if we are to draw generalizable results from the analysis of a sample of privacy policies that includes only P3P-enabled websites, we must first consider whether the policies of P3P-enabled sites are representative of all website privacy policies. We used our set of 67 APPEL rulesets to compare the P3P policies of 20 P3P-enabled sites from the Popular list to the coded policies of 48 sites without P3P policies from the Popular list to determine whether there are significant differences between these two groups.[12] We also compared coded policies of 68 sites on the Random list that did not use P3P with 7,741 P3P-enabled sites found in the Privacy Finder cache.

8.1 Popular Sites

Overall, we found that the policies of popular P3P-enabled sites look quite similar to the policies of popular sites that do not have P3P. A chi-square test indicated no significant difference in the types of data collected, dispute resolution procedures, purpose for data collection, and data retention policies. There were also no significant differences in the percentage of sites matching each of the Privacy Finder privacy levels. However, there were two areas that yielded significantly different results: data recipients and data access. The most significant difference that we found was with regard to data recipients ($\chi^2 = 27.66$, $df = 4$, $p < 0.001$). Table 12 highlights these differences. We see that the sites without P3P are more likely to share data with unrelated third parties, third party affiliates, as well as the general public.

	P3P	Non-P3P
`<delivery/>`	20%	19%
`<other-recipient/>`	25%	0%
`<public/>`	10%	33%
`<same/>`	25%	85%
`<unrelated/>`	10%	38%

Table 12
Comparison of the data recipients policies between P3P-enabled sites and sites without P3P on the Popular list. These differences were significant ($\chi^2 = 27.66$, $df = 4$, $p < 0.001$).

We examined policy differences that concerned individual access to one's personal data, and found a similar trend. These differences can be seen in Table 13. The websites without P3P are more likely to give access to information beyond identified contact information. However, the websites with P3P are less likely to collect personally identified information to begin with.

[12] There are 21 P3P-enabled sites on the Popular list, but one of the P3P policies contained a critical error and thus could not be evaluated. This critical error did not exist when we performed the experiment in Sections 6.3 and 7.2.

	P3P	Non-P3P
`<all/>`	5%	0%
`<contact-and-other/>`	65%	79%
`<ident-contact/>`	20%	6%
`<none/>`	0%	6%
`<nonident/>`	15%	2%
`<other-ident/>`	0%	0%

Table 13
Comparison of the data access policies between P3P-enabled sites and sites without P3P on the Popular list. These differences were significant ($\chi^2 = 9.99$, $df = 4$, $p < 0.05$).

8.2 Random Sites

We also examined less popular sites, to verify that the similarities between the privacy policies of P3P-enabled sites and non-P3P-enabled sites are not unique to popular sites. We followed a very similar procedure, but this time compared coded natural language privacy policies from 63 sites on the Random list with 7,741 P3P-enabled sites found in Privacy Finder's cache.

Because of the large number of ambiguities in these natural language policies, the coded policies contained many "unclears." Thus we were unable to do a direct comparison across all P3P elements. Therefore, we focused our analysis on the three Privacy Finder privacy levels. We evaluated all of the sites against these rulesets and then performed a chi-square test to determine significance. Table 14 shows the aggregate data from these comparisons.

	P3P	Non-P3P
`High`	21%	37%
`Medium`	56%	60%
`Low`	66%	94%

Table 14
Comparison of the preset privacy levels between randomly selected P3P-enabled sites and sites without P3P policies. These differences were significant ($\chi^2 = 10.50$, $df = 2$, $p < 0.01$).

We found that there was a significant difference between data sets ($\chi^2 = 10.50$, $df = 2$, $p < 0.01$). We performed a z-test for proportions on the individual privacy levels independently and found there to be a significant difference ($p < 0.05$) at the "low"' and "high" settings, but not at the "medium" privacy setting. The sites without P3P were less likely to trigger warnings at all three privacy levels than the P3P-enabled sites. It is likely that this is at least partially an artifact of the large number of “unclears” in these policies. When we encounter an “unclear” element we give the website the benefit of the doubt and assume it does not collect that particular data element or engage in that particular practice. Thus we err on the side of triggering fewer privacy warnings.

9 Conclusion

There is increasing media attention paid to identity theft, data aggregation, and online privacy in general. However most users will not take the time to read the privacy policies they encounter, and those who do may be unable to fully understand them. The P3P specification was created to address this problem. For P3P user agents and services such as Privacy Finder to be useful, P3P policies need to be available on a large number of websites. While roughly ten percent of all sites studied have deployed P3P, more than twice as many e-commerce sites have deployed P3P. In addition, P3P deployment rates are highest among the most popular websites and those most frequently returned in search results. We have also shown that P3P adoption is increasing, although at a slow pace in most sectors. However, deployment of P3P by even a few additional very popular sites could substantially increase the frequency with which P3P-enabled hits are returned in search results.

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.[13] 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.[14] Of these, 9,645 (39%) have P3P policies. On the other hand, examining the “.mil” websites 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 websites represent by far the largest sector to adopt P3P.

While a ten percent adoption rate after four years might seem paltry, many

[13] P.L. 107-347.

[14] 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 websites. There are also federal government websites that do not have a .gov domain name. Thus, we can only make a rough estimate about the rate of government P3P adoption.

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 [37]. 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) [38]. Additionally, CSS 2.0 became a W3C standard in 1998, yet as of 2006, there are no web browsers that fully support it [39]. Yet many websites now use some version of CSS.

Beyond examining P3P deployment rates, we have also examined privacy policy trends. We analyzed the content of P3P privacy policies from a variety of industries and found that privacy practices vary significantly across different types of websites. P3P facilitates the collection of data on a much larger number of privacy policies than would be otherwise feasible.

Additionally, we checked P3P policies for syntactic errors and examined their accuracy. We found large numbers of syntactic errors as well as numerous discrepancies between P3P policies and their natural language counterparts. Most of the syntactic errors were not critical to policy evaluation, and many of the discrepancies did not impact Privacy Finder's evaluation of a policy. However, these errors do raise concerns about the reliability of both P3P policies and natural language privacy policies and highlight the need for better tools for authoring and managing both natural language and computer-readable privacy policies.

Finally, we explored the differences and similarities in privacy policies between sites that choose to post P3P policies and those that do not. Among the most popular websites, there is little difference between the privacy practices of sites with P3P policies and sites without P3P policies. We found some significant differences when we examined random sites; however, the large numbers of ambiguities in the natural language privacy policies that we coded limit our ability to draw conclusions from this analysis.

10 Acknowledgments

This work was supported in part by the IBM Open Collaborative Research Initiative and 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, 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.

References

[1] CBS News, Poll: Privacy Rights Under Attack (October 2, 2005).
`http://www.cbsnews.com/stories/2005/09/30/opinion/polls/main894733.shtml`

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

[3] Privacy Leadership Initiative, Privacy notices research final results (December 2001).
`http://www.ftc.gov/bcp/workshops/glb/supporting/harris%20results.pdf`

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

[5] M. J. Culnan, 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`

[6] T. Vila, R. Greenstadt, D. Molnar, Why we can't be bothered to read privacy policies: Models of privacy economics as a lemons market, in: Proceedings of the 2003 International Conference on Electronic Commerce (ICEC 2003), Pittsburgh, PA, 2003, pp. 403–407.

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

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

[9] L. Cranor, B. Dobbs, S. Egelman, G. Hogben, J. Humphrey, M. Schunter, D. A. Stampley, R. Wenning, The Platform for Privacy Preferences 1.1 (P3P1.1) Specification (November 2006).
`http://www.w3.org/TR/P3P11/`

[10] 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) (2002) 276–306.

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

[12] L. Cranor, D. Weitzner, Summary Report - W3C Workshop on the Future of P3P, Tech. rep., World Wide Web Consortium (November 2002).
http://www.w3.org/2002/12/18-p3p-workshop-report.html

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

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

[15] L. F. Cranor, M. Arjula, P. Guduru, Use of A P3P User Agent by Early Adopters, in: WPES '02: Proceedings of the 2002 ACM workshop on Privacy in the Electronic Society, ACM Press, New York, NY, USA, 2002, pp. 1–10.
http://doi.acm.org/10.1145/644527.644528

[16] L. F. Cranor, P. Guduru, M. Arjula, User Interface for Privacy Agents, ACM Transactions on Computer-Human Interaction 13 (2).
http://portal.acm.org/citation.cfm?doid=1165734.1165735

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

[18] D. Fellows, Search Engine Users.
http://www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf

[19] E. Burns, Search Increased in August (October 7, 2005).
http://www.clickz.com/stats/sectors/search_tools/article.php/3554731

[20] L. F. Cranor, S. Byers, D. Kormann, 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.

[21] J. Gideon, S. Egelman, L. Cranor, A. Acquisti, Power Strips, Prophylactics, and Privacy, Oh My!, in: Proceedings of the 2006 Symposium on Usable Privacy and Security, 12-14, July 2006.
http://cups.cs.cmu.edu/soups/2006/proceedings/p133_gideon.pdf

[22] G. R. Milne, M. J. Culnan, Using the content of online privacy notices to inform public policy: A longitudinal analysis of the 1998-2002 U.S. web surveys, The Information Society 18 (5) (2002) 345–359.

[23] S. Beitzel, E. Jensen, D. Lewis, A. Chowdhury, A. Kolcz, O. Frieder, Improving automatic query classification via semi-supervised learning, in: Proceedings of The Fifth IEEE International Conference on Data Mining, New Orleans, Louisiana, U.S.A., 2005.

[24] Google, Inc., Froogle (2005).
http://froogle.google.com/

[25] P. Beatty, I. Reay, S. Dick, J. Miller, P3P Adoption on E-Commerce Web Sites: A Survey and Analysis, IEEE Internet Computing 11 (2) (2007) 65–71.
http://doi.ieeecomputersociety.org/10.1109/MIC.2007.45

[26] C. Jensen, C. Sarkar, C. Jensen, C. Potts, Tracking Website Data-Collection and Privacy Practices with the iWatch Web Crawler, in: Proceedings of the 2007 Symposium On Usable Privacy and Security (SOUPS), ACM Press, Pittsburgh, PA, 2007.

[27] S. Brin, 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

[28] L. Cranor, A. McDonald, S. Egelman, S. Sheng, 2006 Privacy Policy Trends Report, Tech. rep., Carnegie Mellon CyLab (January 31, 2007).

[29] I. K. Reay, P. Beatty, S. Dick, J. Miller, A Survey and Analysis of the P3P Protocol's Agents, Adoption, Maintenance, and Future, IEEE Transactions on Dependable and Secure Computing 4 (2).

[30] S. Byers, L. F. Cranor, 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

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

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

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

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

[35] S. Fortunato, A. Flammini, F. Menczer, A. Vespignani, The egalitarian effect of search engines, Tech. rep., arXiv.org e-Print Archive (2005).
http://arxiv.org/pdf/cs.CY/0511005

[36] L. F. Cranor, Web Privacy with P3P, O'Reilly and Associates, 2002.

[37] H. W. Lie, B. Bos, Cascading Style Sheets, Level 1 (December 1996).
http://www.w3.org/TR/CSS1

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

[39] B. Bos, H. W. Lie, C. Lilley, I. Jacobs, Cascading Style Sheets, Level 2, CSS2 Specification (May 1998).
`http://www.w3.org/TR/REC-CSS2/`