CyLab Privacy Interest Group

2006 Privacy Policy Trends Report

January 31, 2007

Lorrie Faith Cranor, Aleecia M. McDonald,
Serge Egelman, and Steve Sheng

CyLab
Carnegie Mellon University
Pittsburgh, PA 15213

Table of Contents

1 Executive Summary

In this report we examine the state of online privacy at the end of 2006 through the lens of website privacy policies. We look at three main areas: privacy practices of the most popular websites as compared with a random sample of websites that post privacy policies, privacy policies of websites in the US financial industry, and trends in the adoption of the Platform for Privacy Preferences (P3P).

In our first section, *Comparison of Popular Websites to Random Websites,* we contrast the privacy practices of the most-visited websites to the rest of the web. We see how privacy protections differ between the most popular websites and a random selection of websites. Popular sites are still more likely to provide privacy policies than random sites. However, while the percentage of random sites with privacy policies has improved from 77% in 2001 to 88% in 2006, popular sites fell slightly from 99% in 2001 to 96% in 2006. At the highest level, the most popular sites collect more data and share it widely. However, the randomly selected sites provide fewer ways for customers to contest errors. We also demonstrate that privacy policies still require a college education to understand.

In our second section, *Focus on Financial Industry,* we take an in-depth look at the effect of the Gramm-Leach-Bliley Act (GLB) on the financial industry. We find the information available to consumers about financial institution privacy practices is more concrete, with fewer uncertainties since GLB went into effect. Unfortunately, those practices have not improved – and data sharing is even more widespread today than before the law was enacted. We also find that while privacy policies still require a college education to understand, institutions switched from using an industry standard privacy policy to a sample FCC privacy policy. This suggests future outcomes could be improved by providing better sample policies.

In our third section, *Platform for Privacy Preferences,* we look at P3P-enabled websites. We find the most popular sites are more likely to have P3P than less popular websites and that P3P has world-wide acceptance. P3P-enabled websites in the European Union are more privacy protective than non-EU P3P-enabled websites. We discuss differences between privacy practices by industry segment (shopping, government, news and media, computers, banking, business to business, adult, blogs, and education). We show that P3P deployment continues to increase. Finally, we present an analysis of errors in P3P policies. While we found 73% of P3P policies have errors, only 5% of those are critical errors.

2 Introduction

As many studies have shown, Internet users are concerned about their privacy.[1] After the California Security Breach Information Act of 2003 (S.B. 1386) became law, several large corporations announced the extent to which they lost control of customer's information. The number of announcements, the stature of the companies, and the staggering numbers – upwards of 50 million accounts put at risk – combined to make 2005 a bad year for privacy.[2] Worse, most experts agreed there was nothing special about 2005, just increased transparency that highlighted long-standing problems.[3]

Privacy breaches continued to grab headlines in 2006. Highlights include:

- Veterans Affairs lost personal information for about 26.5 million people when an employee took a laptop home, only to lose it in a burglary.[4]
- An employee of the St. Louis Red Cross had access to personal information for one million donors, including their social security numbers, and committed multiple cases of identity theft.[5]
- Hewlett Packard's board of directors used pretexting, email beacons, and read instant messages to attempt to find who was leaking information to the press. This resulted in Congressional investigation and several high-level resignations from Hewlett Packard.[6]
- Approximately 19,000 purchasing records were stolen from one of AT&T's vendors. The theft was followed by a targeted phishing attempt which was designed to collect additional credit card information.[7]
- AOL provided about 20 million search queries to researchers, but did not take adequate measures to ensure the data was fully anonymous. Two employees were fired, and AOL's Chief Technical Officer resigned.[8]

As this small sample shows, privacy breaches cause harm to the companies responsible, resulting in public relations nightmares, dismissal of key personnel, loss of customer trust, and in some cases Congressional oversight. For consumers, harms include identity theft and trouble with credit records. The Hewlett Packard and AT&T vendor examples are particularly interesting because they highlight a growing trend: targeted privacy invasions where the subject is specifically selected.[9]

Policy responses have varied in 2006. While some legislation has been introduced to protect privacy, other legislation could result in further privacy erosion. For example, Representative Edward Markey introduced the Eliminate Warehousing of Consumer Internet Data Act to mandate companies discard data they no longer need, but Representatives Barton and DeGette suggested mandatory data retention that would require ISPs (Internet Service Providers) retain customer information for at least a year to facilitate child pornography investigations.[10]

At the close of 2005, Microsoft endorsed proposed privacy legislation, which is a change from the typical model of industry self-regulation.[11] Microsoft's announcement was praised by the Center for Democracy and Technology (CDT), Electronic Privacy Information Center (EPIC), and the American Civil Liberties Union (ACLU.)[12]

A positive privacy development in 2006 was Microsoft's publication of *Privacy Guidelines for Developing Software Products and Services*.[13] While the Microsoft document still leaves controversial practices without guidelines, such as adware and location-based services, it is a serious step forward. In particular, Microsoft notes in the second sentence that many people consider privacy a human right, operates within the Organisation for Economic Co-operation and Development (OECD) Fair Information Practices, calls for increased transparency, and states "publishing a P3P policy is recommended" for websites.[14] (See our "Platform for Privacy Preferences" section on page 34.) Moreover, the document moves beyond theory and into the realm of practical examples. It shows software developers how to incorporate appropriate privacy protections in nine common use case scenarios.

Interest in privacy remains high and the United States continues to take a largely hands-off approach for regulating privacy on the Internet. Instead, privacy policies are touted as a crucial component of industry self-regulation.

A 2004 court case held that if customers do not read a privacy policy, their expectations of privacy are low, and the company's privacy policy is unenforceable.[15] However, even though customers may not have standing to win cases without reading privacy policies, the Federal Trade Commission (FTC) can still bring action for fraudulent and deceptive practices.[16] In this way companies are bound to uphold their own privacy policies, although there are few legal requirements as to the content of the policies.

In addition to fraud, the FTC has advanced a legal theory of unfairness and applied it to the practices of companies that do not have privacy policies.[17] The FTC brought action against Vision I LLC d/b/a Cartmanager for practices likely to harm consumers, in response to Cartmanager surreptitiously renting their customer lists.[18] Even though Cartmanager did not contradict their own privacy policy, they still faced FTC action.[19] Neglecting to create a privacy policy, or creating a weak one, is not a reliable defense.

With privacy policies a central element of online privacy protections there is a great deal of interest in the following questions:

- How prevalent are privacy policies?
- Are privacy policies usable? That is, can customers read and understand them?
- Are companies improving or stagnating on the use of privacy policies?
- What privacy protections do policies contain?
- Do industries differ in their privacy practices?

To answer these questions we turn to the privacy policies themselves. We are particularly interested in noting privacy trends from year to year so we can determine the direction and rate of change.

2.1 Methodology

Human-readable privacy policies do not lend themselves to statistical comparison and analysis. We hired two students to read privacy policies and code their contents by answering a series of multiple choice questions about each policy. We used software we developed to present a policy, data entry form, and code book to facilitate data collection and data entry. Policies generally took 20 to 40 minutes to code. When policies did not explicitly mention a particular data practice and it was not possible to infer it through a cursory examination of forms on the website, that practice was coded as "unclear."

Our software converted the multiple-choice forms completed by the students into computer-readable P3P format, slightly altered to allow for "unclear" codes (see our "Platform for Privacy Preferences" section for more information about P3P). As we discuss below, among other things P3P provides a taxonomy of privacy policy elements. By using P3P as a common representation we were able to compare all of the natural language privacy policies that we examined. A doctoral candidate experienced with P3P coding checked policies for accuracy.

In addition, many websites implement P3P and publish their privacy policies directly in P3P form. We used our software to analyze the privacy P3P policies of approximately 8,000 P3P-enabled websites.

Finally, in some cases we had both the P3P policies from a website as well as our translation of their human readable policies into P3P format. We were able to contrast the P3P they provided with the P3P we coded, to see if their human readable policies match their P3P policies.

2.2 Data Sets

We used or created the following data sets:

1. **AOL search most-clicked domains.** 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, described below.
2. **Most popular websites.** We selected the most popular websites from the AOL search data. In order to make our results comparable to other studies in the Milne and Culnan study[20], we further refined the list by removing all 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. See Appendix A for the list of websites we used.
3. **Random websites.** We again used the AOL search data and omitted sites other than .coms, pornographic content, and sites targeted to children. Of the top 12,000 most popular websites, we selected 100 at random. We limited the sample frame to the top 12,000 in order to be comparable to other studies. Of the 100 websites on our Random list, 78 had human-readable privacy policies and 9 had both human-readable policies and P3P policies. See Appendix B for the list of websites we used.
4. **Financial websites.** We compiled a list of the top 10 U.S banks with a significant online presence from 1999 to 2005, which we extracted from the list of top 500 depository institutions (SIC code 602).[21] We also generated a list of 30 randomly selected U.S. banks with a significant online presence. In addition, we created a dataset of the top 10 credit card issuers. Finally, we selected 10 retail websites at random to serve as a control group.

5. **Longitudinal financial websites.** We collected privacy policies from 60 U.S. companies (banks, credit card, random retail as mentioned above.) For each company, we collected their privacy policies once a year from 1999 to 2005. 2005 policies were collected from the companies' website directly. Policies from 1999 to 2004 were collected from the Internet Archive.[22]
6. **P3P-enabled websites.** We used a list of P3P-enabled websites discovered by the Privacy Finder[23] search engine. The search engine maintains a cache of every P3P-enabled website that is returned from a user query. As of December of 2006, Privacy Finder's cache contained over 150,000 websites, 9,408 of which were P3P-enabled.
7. **Industry segmented websites.** We were able to define categories for a portion of the websites in the Privacy Finder cache using information from the Yahoo! Directory. Using a custom script, we were able to categorize 16,919 sites that were in our cache (about 11% of our cache). Of these, 1,181 were P3P-enabled (7%). Due to the large number of categories yielded in this fashion, we decided to only analyze the most popular categories: "shopping," "government," "news and media," "computers," "banking," "B2B (business to business)," "adult," "blogs," and "education."
8. **Typical search terms.** 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.
9. **Froogle search terms.** We 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 generally are going to be indicative of e-commerce. Using a Perl script, we screen-scraped these search terms from Froogle. We collected 940 unique terms in this manner.

3 Comparison of Popular Websites to Random Websites

We coded privacy policies from the top 75 websites ("Popular") and a random sample of 100 websites ("Random") into standard P3P format. We checked the policies against the privacy settings used by the Privacy Bird P3P user agent. We also analyzed the policies against sixty-two rule sets developed for a 2003 study.[25] We follow a similar approach as used in that study, and compare our results to that study in section Analysis of Privacy Protections on page 11.

3.1 Privacy Bird Settings

Privacy Bird[26] is an Internet Explorer browser helper object designed to help users determine whether a website conforms to the user's notion of acceptable privacy protections. A small bird icon appears in the browser title bar. If the privacy policy of the site meets or exceeds what the user indicates as acceptable privacy protections, the bird is green. If the site has privacy policies that do not meet the user's privacy threshold, the bird turns red. If there is no P3P policy at that site the bird turns yellow.

Figure 1: Privacy Bird red, green, and yellow icons.

Users are able to select their own privacy settings based on their personal privacy preferences. In addition, Privacy Bird is pre-configured with three standard settings for "low," "medium," and "high" privacy. Figure 2 includes the description of each setting from the 2003 study.[27]

Privacy Bird Settings

Low. Trigger a red bird at sites that collect health or medical information and share it with other companies or use it for analysis, marketing, or to make decisions that may affect what content or ads the user sees. Also trigger a red bird at sites that engage in marketing but do not provide a way to opt-out.

Medium. Same as low, plus trigger a red bird at sites that share personally identifiable information, financial information, or purchase information with other companies. Also trigger a red bird at sites that collect personally identified data but provide no access provisions.

High. Same as medium, plus trigger a red bird at sites that share any personal information (including non-identified information) with other companies or use it to determine the user's habits, interests, or other characteristics. Also trigger a red bird at sites that may contact users for marketing or use financial or purchase information for analysis, marketing, or to make decisions that may affect what content or ads the user sees.

For all three settings, a site is classified as not sharing data if it shares data only with agents that use it only to complete the transaction for which it was provided or with delivery companies (which may have unknown data practices). In addition, the Privacy Bird settings classify a site as not sharing data if data sharing occurs only under an opt-in policy.

For these three settings data from the following P3P categories are considered personally identifiable information: physical contact information, online contact information, and government issued identifiers.

Figure 2: Privacy Bird Settings

It may be easier to understand the Privacy Bird settings by way of concrete example, as in Figure 3.[28]

Warn when...	Low	Medium	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		

Figure 3: Privacy Bird Warnings

We used the Privacy Bird settings to analyze the privacy protections offered by Popular and Random websites, based on coding their human-readable policies into P3P. The percentages below indicate how frequently a user would get a red bird warning them not to proceed.

	Popular	Random
Low Setting	14.1%	30.4%
Medium Setting	42.2%	52.2%
High Setting	68.8%	82.6%

Table 1: Privacy Bird Evaluation of Websites: Percentage of Sites Receiving "Red Birds" Under High, Medium, & Low Settings

Overall, the Popular websites are more privacy protective than the Random websites. In some cases we were not able to determine if a site gathered particular data or engaged in a particular practice, and thus coded it as "unclear." Some practices were never unclear, while others were unclear as much as 76% of the time. In our analysis, if a site did not explicitly mention a particular practice or was unclear about it, we assumed they did not engage in that practice. Therefore, we may be undercounting the number of sites that should receive "red bird" warnings at each setting.

3.2 Previous Studies

The 2006 Privacy Policy Trends Report is our first in a series of annual summaries. We designed our research to allow us to compare our results to prior studies.

The Federal Trade Commission (FTC) started annual privacy policy surveys starting in 1998. Privacy policies surveys have been an important tool for informing public policy debate on protecting privacy. For example, the FTC conducted "web sweeps" in 1998 and 2000 by surveying privacy policies posted online.[29] The Federal Reserve System and several other agencies surveyed privacy practices posted by financial institutions in 1999.[30] Finally, Milne and Culnan's work combines previous surveys. Their longitudinal analysis of 1998 – 2001 privacy policies takes into account issues from different protocols used over the years, and normalizes them to allow direct comparison between the surveys.[31]

The rest of this section contrasts our findings in 2006 to the normalized findings from the Milne and Culnan study. Note that while we followed similar protocols our results have not been normalized and therefore may not be directly comparable. We offer this comparison to give an idea of general trends over time.

Note that because the web surveys collected different data over the years, in some cases data is unavailable.

	1999	**2000**	**2001**	**2006**
Number of sites surveyed	*n* = 286	*n* = 281	*n* = 223	*n* = 100
Privacy policy	48.3%	65.8%	76.7%	88%
Provides notice about what personal information is collected	49.7%	71.2%	73.5%	100%
Provides notice about disclosure to third parties	40.6%	N/A	N/A	83%
Provides access	27.6%	21.4%	N/A	94%

Table 2: Comparison of privacy practices between random sample groups

	1998	1999	2000	2001	2006
Number of sites surveyed	*n* = 105	*n* = 91	*n* = 87	*n* = 71	*n* = 75
Privacy policy	44.8%	84.6%	96.6%	98.6%	96%
Provides notice about what personal information is collected	N/A	73.6%	90.8%	95.8%	100%
Provides notice about disclosure to third parties	58.1%	71.4%	N/A	N/A	80%
Provides access	26.7%	41.8%	49.4%	N/A	95%

Table 3: Comparison of privacy practices between popular sample groups

3.3 Analysis of Privacy Protections

What practices differ between Popular and Random websites to cause such a difference in privacy protections? The remainder of this section discusses the privacy practices of the Popular and Random websites in more detail.

P3P separates data gathering and use into several different categories, which we discuss below. The P3P elements are defined and discussed in greater detail in the P3P specification.[32] For example, when "computer information" is collected, that means "information about the computer you are using, such as its hardware, software, or Internet address."[33] Please see Appendix C: P3P Elements on page 82 for the recommended human-readable descriptions of P3P elements.

Because we coded P3P policies from human readable privacy policies, we could not always tell what practices a website follows. For example, if a website does not mention the common practice of collecting information from server logs, does that mean they don't collect it or simply omitted mention? We added a new category of "unclear" to capture data on the lack of information provided in human readable policies. It may help to think of the data reported without the unclear category as the lower bound, and with unclear as the upper bound.

3.3.1 Types of Data Collected

One of the most fool-proof ways to protect data, and therefore privacy, is to simply never collect it in the first place. Of course this is at odds with many corporate and consumer needs: a company cannot ship a book to your home without your address.

Most sites (over 82% for Popular, over 69% for Random) collect computer, state management, and navigation information. Any site that does not deliberately turn off web logging is likely to collect that data.

Most sites ($\geq$ 78% Popular, $\geq$ 55% Random) also collect online contact information, physical contact information, and unique identifiers. Ecommerce sites need such information to ship products. Informational sites like news papers do not even need unique identifiers to deliver the news, but many require unique login accounts.

Few sites ($\leq$ 40% Popular, $\leq$ 17% Random) collect health, political, or location-based data.

Popular sites are, generally speaking, more likely than Random sites to collect more types of data. However, they are also more likely to disclose which types of data they collect.

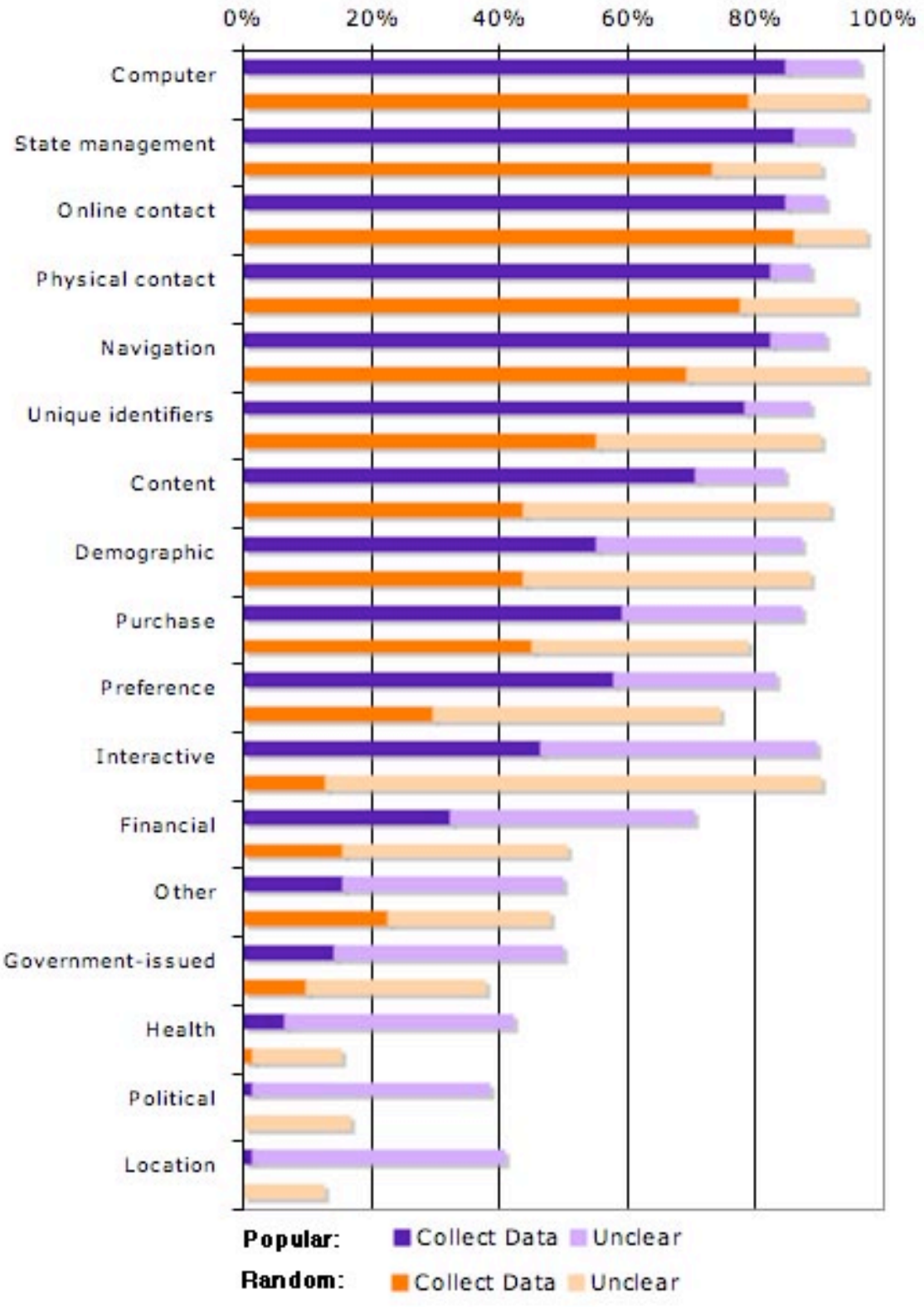

Figure 4: Percentage of Sites Collecting Each P3P Data Type

3.3.2 Data Use

Data use captures how sites use data above and beyond the expected purposes it was supplied for.[34] Sites frequently use data for system administration and research and development. Pseudonymous use may be tied to marketing, for example using a cookie to track a customer's interests over time and serving different ads as a result. This does not require any knowledge of the customer's name, address, or other identifiable information.

Nearly all sites use data for system administration and to support the user's current activity. Most sites use data for research and development (97% Popular, 59-92% Random.) This data may be as benign as tracking the number of hits to a website in order to decide when to purchase additional servers. Neither system administration nor research and development usually raise privacy concerns, as long as these uses don't lead to additional secondary data usage.

Many sites use data for pseudonymous analysis (≥ 76% Popular, 27-82% Random) with less use for pseudonymous decision making (10-50% Popular, 21-80% Random). We have a lot of uncertainty in this data, particularly for Random sites, as their privacy policies do not explicitly mention this topic. Most policies that do mention pseudonymous use explain that while they create profiles of their users, generally for advertisement placement and revenues, they do not keep personally identifiable information as part of that data set. This is a compromise that offers some privacy protections while addressing business needs. However, it does not guarantee that profiles cannot be re-identified.

A sizeable portion of sites acknowledge that they use data for future marketing purposes (≥ 47% Popular, ≥ 61% Random.) It is particularly interesting that the Random sites are more likely to engage in this behavior than the Popular sites. While Popular sites are more likely to collect wide varieties of data, they also may be less likely to use that data in ways that surprise their visitors.

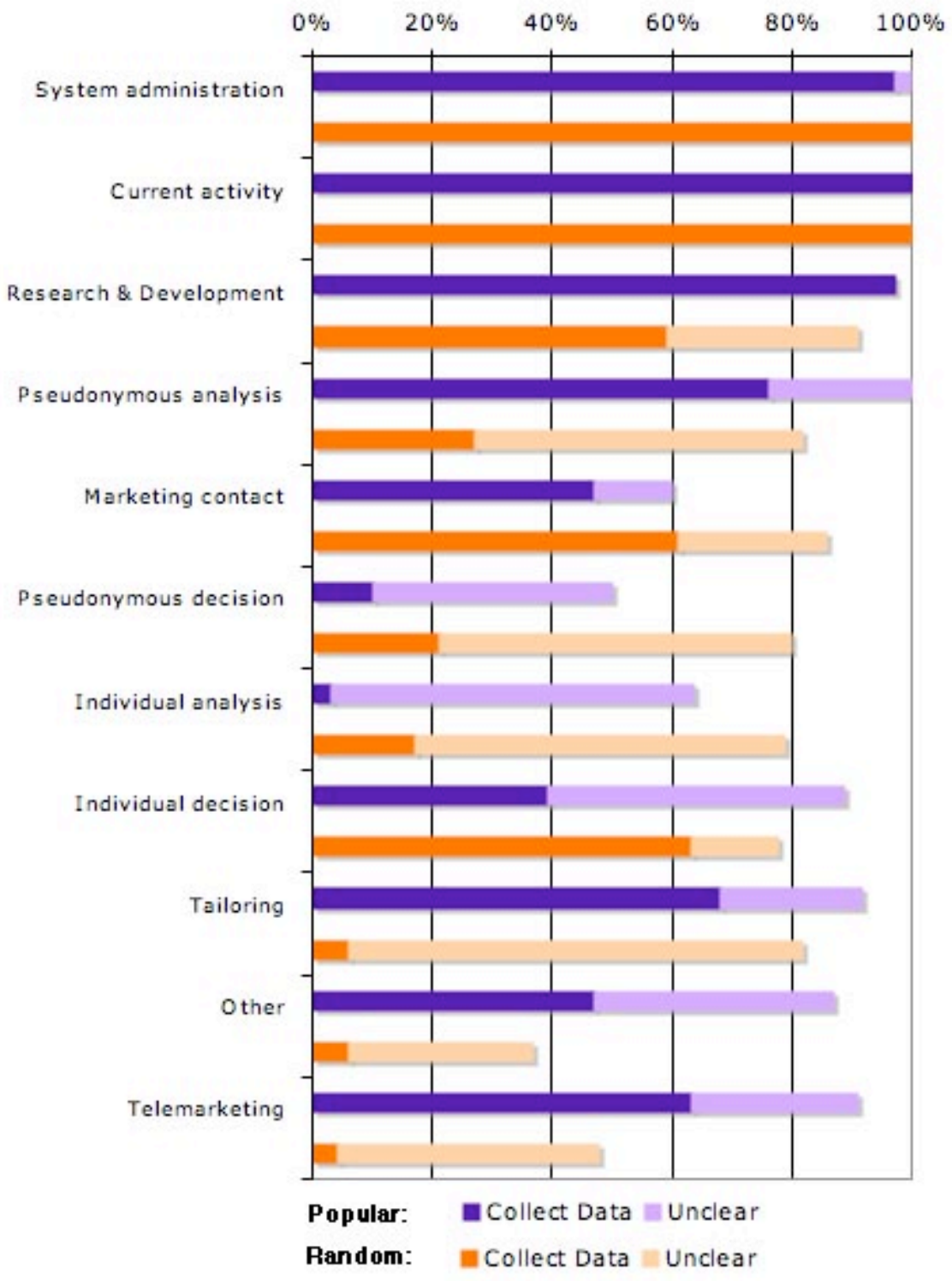

Figure 5: Percentage of Sites Using Data for Each P3P Purpose

3.3.3 Data Recipients and Sharing

Another important facet of privacy is knowing how far data is shared with other companies. Many sites (≥ 54% for Random; ≥ 41% for Popular) share data with other websites that follow the same practices (often affiliates of a parent company, or partners that have entered into a long-term business arrangement.) Fewer sites (≥ 41% Popular, ≥ 15% Random) share data with third parties who may not follow the same practices as the original site. At the far extreme, data is public. One example of a public forum is data sold on a CD. In practice, it is more common that users decide to post information on a message board. This may be a feature that is more common on Popular sites than Random sites, which would account for the difference between them (≥ 38% Popular, 8% Random.)

Delivery services are low in the aggregate (≥ 19% Popular, ≥ 15% Random) and are likely to correlate to the type of website. Ecommerce sites will need to share data with delivery services; government sites probably do not. We expect there are more sites that share data with delivery services but do not explicitly mention it in their privacy statements.

Data sharing is fairly similar between the Popular and Random sites.

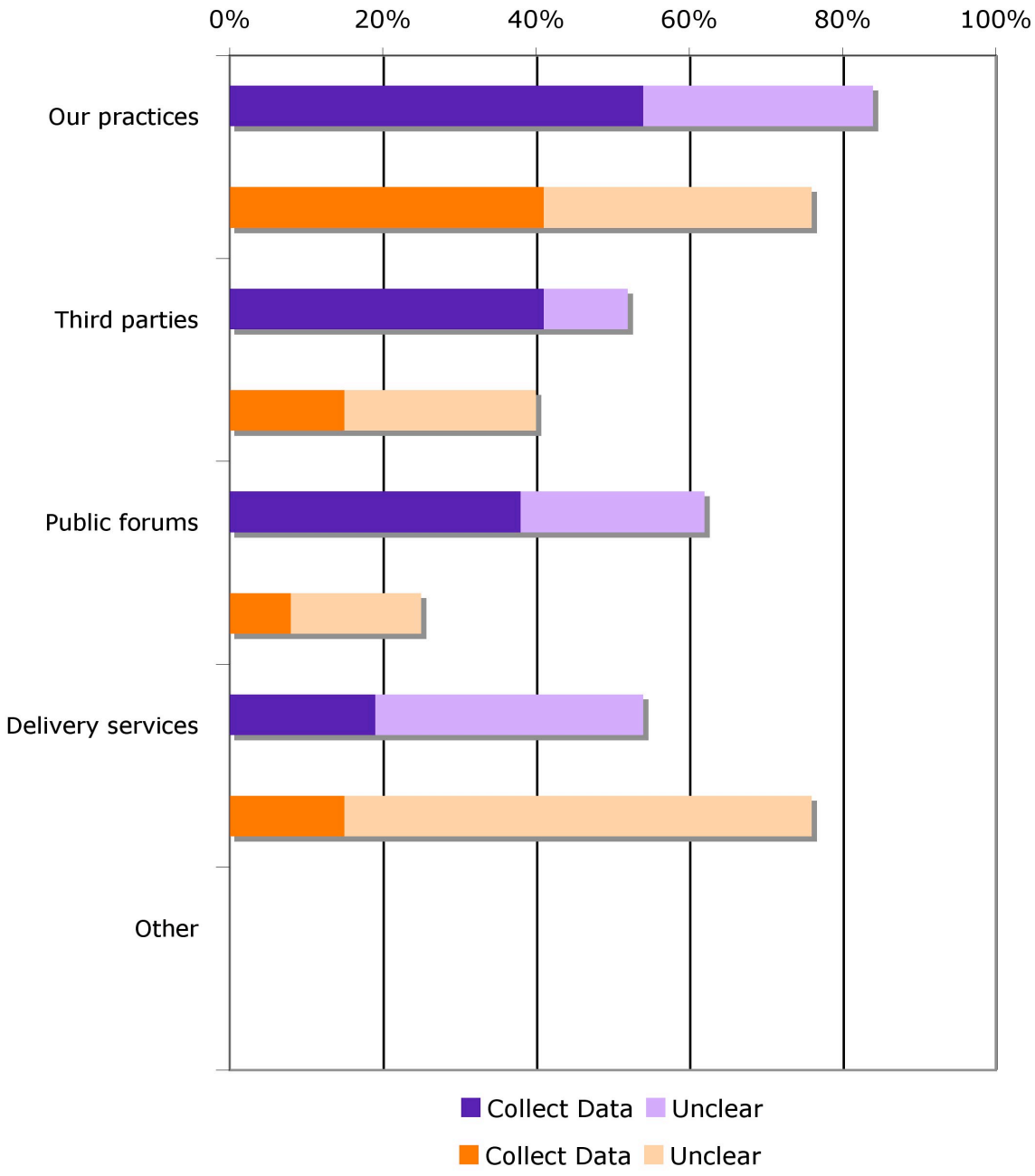

Figure 6: Percentage of Sites Sharing Data with Each Type of P3P Data Recipient

3.3.4 Access Provisions

Fair Information Practices include access as an important component of privacy protections.[35] Access allows people to see what data is held about them. Without access, there is no way to know that information is incorrect.

Websites uniformly give access for customers to change their contact "and other" information (≥ 77% Popular, ≥ 80% Random.) Contact and other means that in addition to contact users have access to some, but not all, information that is collected about them.

Since few sites provide access to all of the data they collect, it is unclear whether the fair information principle of access is being adequately addressed. We cannot tell whether most websites do provide access to the data that individuals are interested in accessing or not. There are also questions about whether complete access would require access to information stored in backup tapes or otherwise not readily accessible in addition to information kept in a company's working database. This is an area policy makers and those involved in industry self-regulation may wish to revisit.[36]

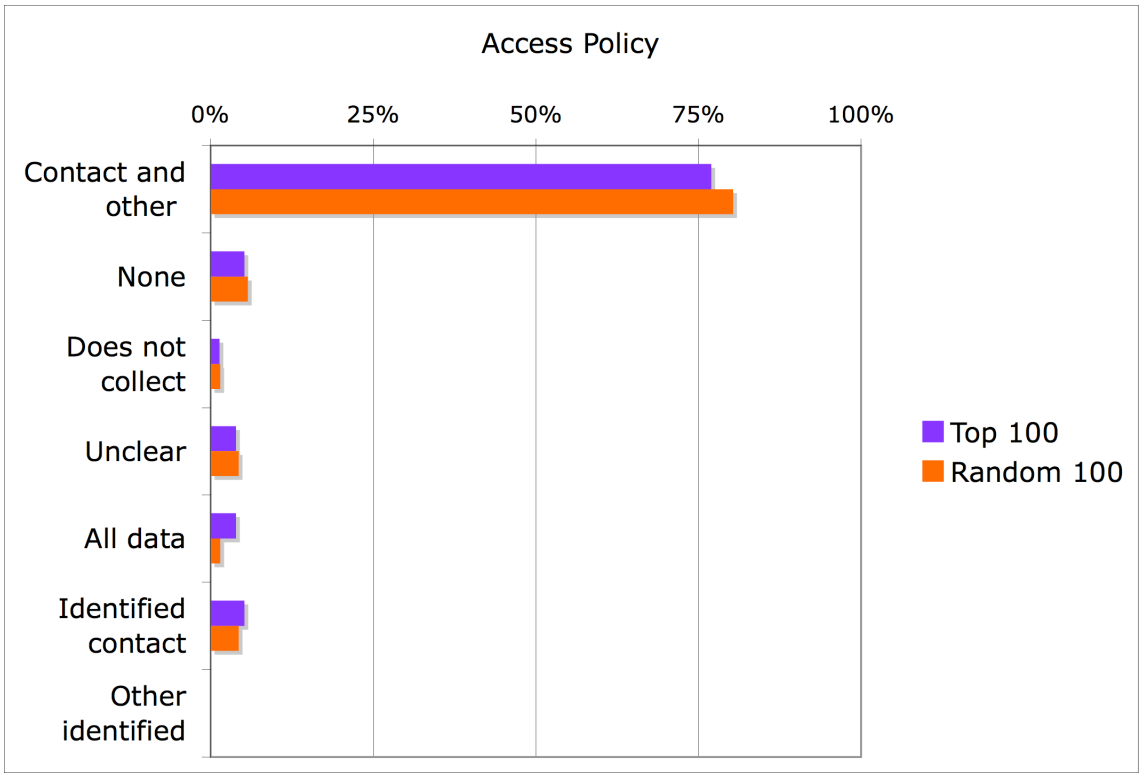

Figure 7: Access Provisions: Percentage of Sites Offering Each Type of P3P Access Provision

3.3.5 Dispute Resolution Options and Remedies

Dispute resolution addresses what recourse customers have when they believe their information is used in a way contrary to the stated practices. Most websites (100 Popular, ≥ 89% Random) direct people to contact their customer service department. In many cases this is phrased as "if you have questions…" rather than "if you have problems…" but the contact information tends to be easy to find.

Independent organizations include privacy seal programs such as TRUSTe[37] and BBBOnline[38]. Comparatively few websites (19% Popular, 20% Random) work with independent organizations. It is interesting that the Popular and Random sites are similar in their adoption rate, particularly because these programs require high institutional effort and resources.

Nearly no sites suggested law or courts as a potential recourse. Even when data is protected by industry-specific laws, companies do not volunteer that information.

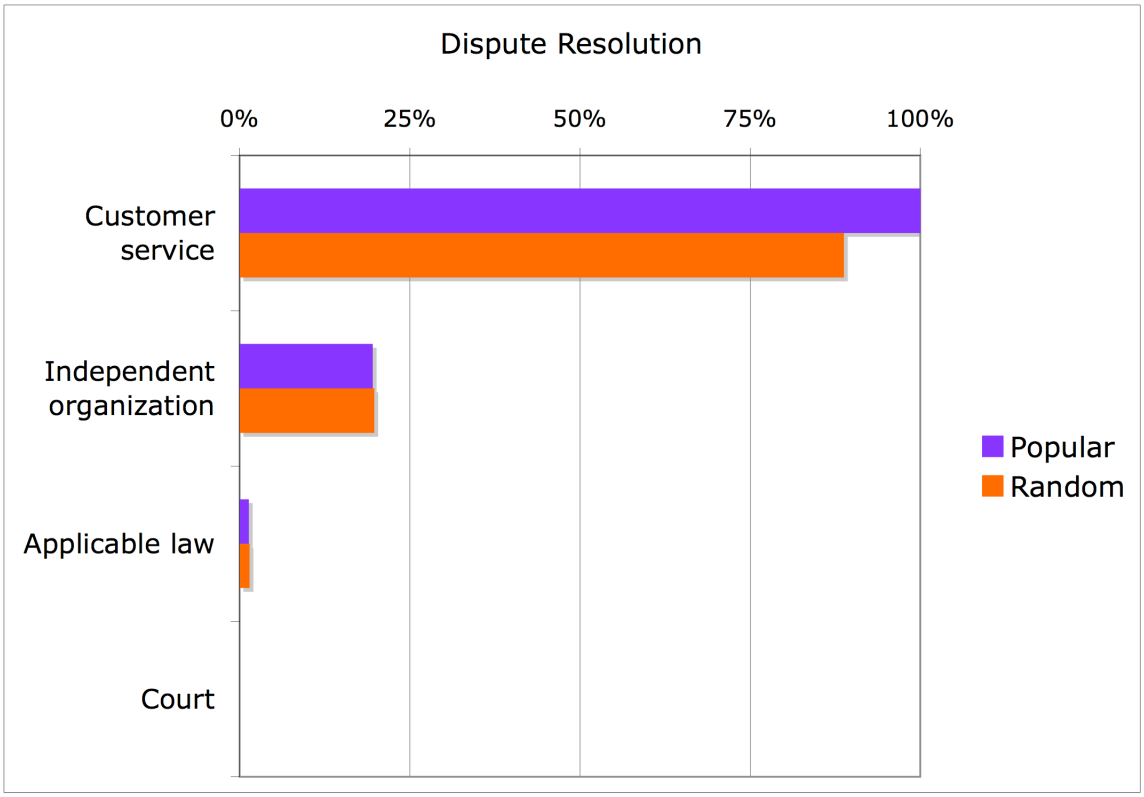

Figure 8: Dispute Resolution Options and Remedies Offered: Percentage of Sites Offering Each Type of P3P Dispute Resolution Option

3.3.6 Data Retention Policies

Privacy is threatened by the indefinite retention of information. For instance, while one record of what you purchased does not indicate much about you, a record of all of your purchases over the past ten years may provide a very detailed portrait of your lifestyle.

Websites are largely silent about their retention policies, which results in a great deal of uncertainty about their practices (72% Popular, 70% Random are unclear.)

Of those that provide information, the most common is indefinite retention (17% Popular, 18% Random.) While many cites explicitly warn customers that data may be turned over to law enforcement, data retention policies do not appear to be driven by laws.

This is another area that may benefit from scrutiny.

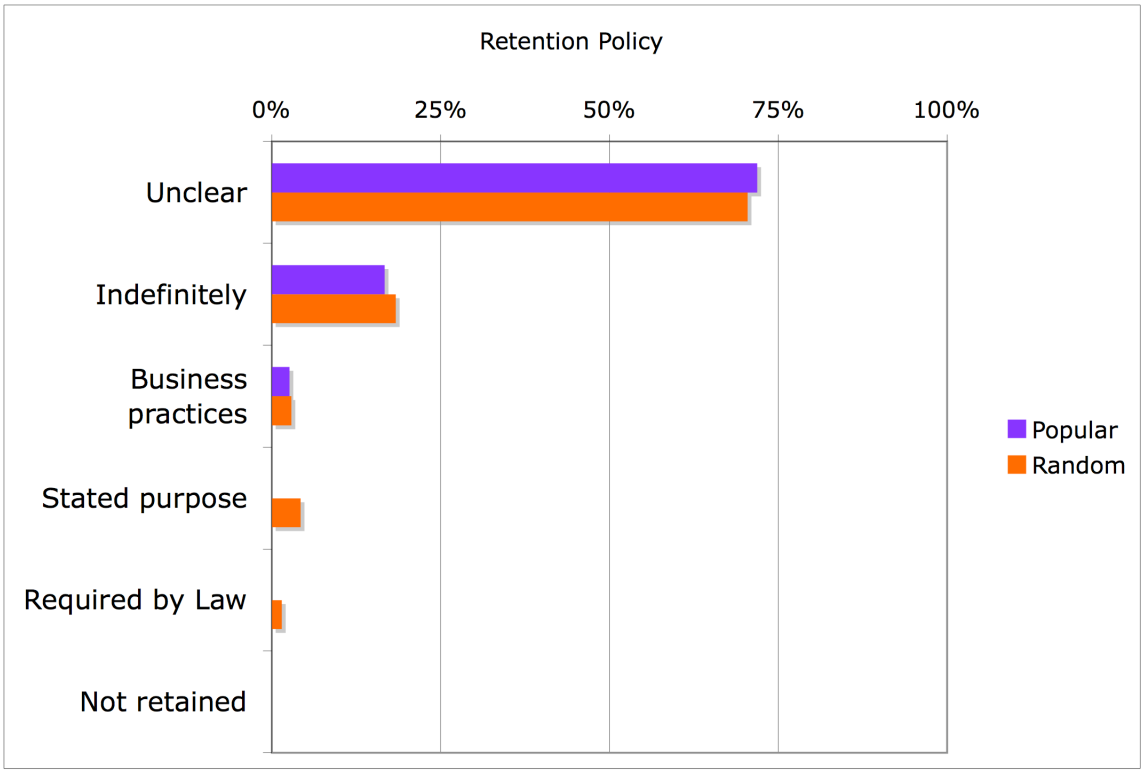

Figure 9: Data Retention Policies: Percentage of Sites Offering Each Type of P3P Data Retention Policy

3.3.7 Differences in Privacy Protections

In general, the Popular and Random policies are fairly similar. However, it is difficult to reach strong conclusions due to the uncertainty we face due to lack of information in human readable policies.

P3P Area	Popular, Percentage Unclear	Random, Percentage Unclear
Type of Data Collected	25% (average)	30% (average)
Data Use	29% (average)	35% (average)
Data Recipients and Sharing	45%	45%
Access Provisions	4%	4%
Data Retention	72%	70%

Table 4: Uncertainty in Human Readable Policies. Type of Data Collected and Data Use are Averaged Over Multiple Sub-categories.

This high level of uncertainty is, itself, an interesting result. Websites with human readable privacy policies do not offer the level of detail and specificity to enable analysis of many of their practices.

3.4 Analysis of Readability

Anecdotal evidence suggests one of the reasons people are reluctant to read privacy policies is the polices are difficult to understand. Readability analysis supports the view that polices are inaccessible to most of the population.

There are a variety of standard metrics for comparing how easy it is to comprehend a text. In general, readability metrics are ratios based on word length, sentence length, and paragraph length. All other things being equal, we expect that the longer a sentence is, the more complex it is and the harder it is to read. Higher scores indicate more difficult text.

The Kincaid formula[39] is

= [(11.8 * syllables) / number of words] +
[(0.39 * number of words) / number of sentences] – 15.59

The average readability scores for the privacy policies we looked at indicate most people would find them difficult to read. For the sake of contrast, we include typical readability scores for two other forms of communication: press releases from the White House, and New York Times articles.[40]

Metric	Popular	Random	White House Press Release	New York Times Article
Kincaid Score	12.4	12.5	4.1	6.2
Standard Deviation	1.8	2.3		
Minimum Kincaid	7.3	6.8		
Maximum Kincaid	17.9	18.1		

Table 5: Readability Scores for Popular and Random Sites

Certainly there is room for improvement. Unfortunately, corporations have incentive to retain the status quo. Because the FTC uses privacy policies to bring action for fraudulent and deceptive practices, legal departments may be wary of using plain English. Corporations may have fewer liability concerns by using standard, boiler-plate legal language.

3.4.1 Differences in Readability

The Popular and Random are very similar to each other: most policies require a college education to understand.

There is slightly more variability in the Random policies, as seen in both a wider range from minimum to maximum and a slightly higher standard deviation.

4 Focus on Financial Industry

This year we present an in-depth look at the privacy policy trends in the financial industry.

This section of the report includes excerpts from a study published in the Fall 2006 issue of *I/S: A Journal of Law and Policy for the Information Society*. [41]

4.1 Analysis of Privacy Protections

The financial industry is interesting for several reasons. First, consumers who use on-line banking are likely to be particularly concerned with data security and financial privacy. Second, unlike most industries in the United States, the financial industry is subject to Federal legislation that explicitly addresses privacy notices. We looked at how the financial industry responded to the Gramm-Leach-Bliley Act (GLB) legislation.

Overall, we found only modest impacts of GLB on financial institution privacy policies. We found that post-GLB privacy policies are longer, more complete, and more standardized than pre-GLB policies. They are also slightly easier to read, but still require college-level reading skills to comprehend. Most companies we surveyed have adopted policies that do not exceed the mandated requirements and provide a level of privacy lower than that provided by the top retailers we surveyed.

The Financial Modernization Act of 1999, also known as the "Gramm-Leach-Bliley Act" was signed into law on November 12, 1999 and became effective on July 1, 2001. Title V of the GLB Act requires financial institutions to provide an initial "clear and conspicuous" notice of privacy policies and practices to all customers, an annual notice of privacy policies, and an opportunity for consumers to opt out of disclosing protected financial information to nonaffiliated third parties.[42] The FTC's final rule[43] specifies what information should be minimally included in the privacy notices and provides examples of GLB-compliant privacy policies.

However, the GLB has two very important exceptions[44]: first it allows for disclosure to an "affiliate"[45] of the financial institutions without providing notice of the disclosure and opportunity to opt out. Thus personal information may flow freely between affiliated financial institutions. Second, it permits disclosure of nonpublic personal financial information to nonaffiliated third parties that jointly offer marketing with the original institution.

GLB does not preempt other state laws that offer stronger privacy controls. As a result a number of states have taken advantage of this and have enacted their own privacy laws that are stronger than the GLB.[46] For example, the California legislature deemed the GLB protection insufficient and enacted a financial privacy law (California SB 1) that requires companies to give consumers an "opt in" choice before sharing with unaffiliated third parties, and an "opt out" choices before sharing with affiliates. Vermont's Department of Banking, Insurance, Securities, and Health Care Administration also adopted opt-in provisions for information sharing.[47]

Responses to the GLB Act privacy rule have been mixed. Many privacy law scholars were critical of exceptions and the choices offered in the law. For example Neal R. Pandozzi concluded that "Title V (GLB Privacy Section) is riddled with loopholes and exceptions that severely weaken, if not paralyze, the consumers' power to opt out of information sharing between financial institutions and nonaffiliated third parties."[48] Paul M. Schwartz argues that the GLB's promise falls short because the opt-out requirement burdens the consumer. He wrote, "The opt-out rule fails to impose any penalty on the party with superior knowledge — the financial entity — should negotiations over further use and transfer of data fail to occur. … the GLB Act places the burden of bargaining on the less-informed party, the individual consumer."[49]

Peter Swire has criticized some aspects of the GLB Act, but argued in an article titled "The Surprising Virtues of the New Financial Privacy Law," that the law has some "surprising merits." For example, the broad definition of "financial institutions" in the law makes it applicable to many institutions, and the level of detail required by the GLB notice requirement requires financial institutions to examine their data practices and creates more possibilities for accountability. [50]

Was GLB beneficial for customers? We examined privacy policies before and after the GLB went into effect to find out.

4.2 Comparison of Policies Before and After Gramm-Leach-Bliley

What effect has the GLB Act had on financial institution privacy policies and practices? This is the overarching question that our financial industry research seeks to address. There is an implicit assumption that the GLB privacy rule required regulated institutions to change their policies and practices. However, there has been little empirical data collected on the types of changes that were made to comply with the law, the extent of these changes, and the number of institutions impacted.

Because websites are frequently archived by Internet search engines and companies attempting to provide archives of the entire Web, it is now possible to retrieve not only a company's current online privacy policy, but also privacy policies previously posted on their website. Thus a rich set of data on privacy practices is now available going back several years.

We collected privacy policies from 60 US companies with significant online presence:

- The top ten credit card issuers
- The top ten banks
- Thirty randomly selected banks
- A control group of ten retailers

For each company, we collected their privacy policies once a year from 1999 to 2005. Current policies were collected from the companies' website directly. Policies from 1999 to 2004 were collected from the Internet Archive.[51]

4.2.1 Completeness and Standardization Increase After GLB

The privacy policies we examined became longer, more complete, and more standardized after the GLB Act took effect.

We observe that the length of the policies among the random 30 banks increased rapidly from 1999 to 2002, and then leveled off. These changes started before the legislation went to effect (the GLB was passed in November 1999 but did not actually take effect until July 2001), probably due to companies revising their privacy policies so that they would be in compliance by the time the law took effect. It is also possible that due to the increasing attention that data privacy was receiving during that period, companies expanded their privacy policies to promote consumer confidence.

The top 10 banks' policies became more substantial over time, but do not exhibit the same pattern as the random banks' policies. The difference we observe between the two groups is that pre-GLB, the top ten banks privacy policies were already more substantial than the random 30 banks (average of 1504 words as compared to 508 words). Thus, less change was necessary to comply with the law. Interestingly, some top banks with longer policies before GLB shortened their policies after the GLB was enacted. We observed a similar trend in the random group. Credit card companies' policies also became longer over our period of study, increasing from 900 words to 1600 words.

Figure **10** shows the length of policies (measured by the number of words) for the top 10 and random 30 groups of banks from 1999 to 2005.

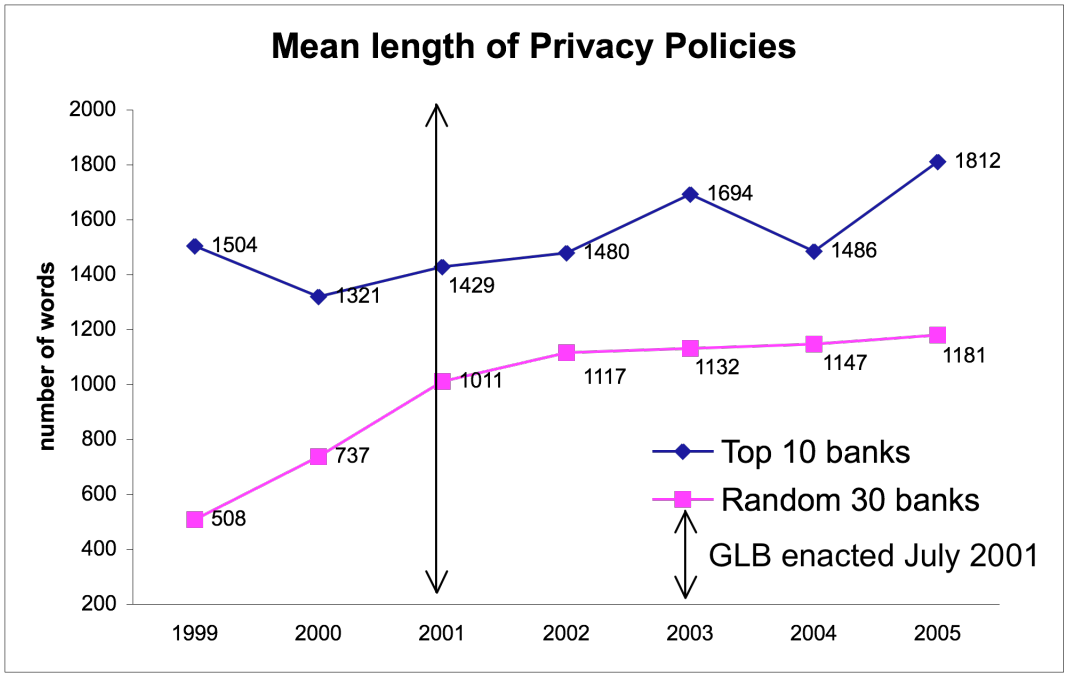

Figure 10: Mean length of privacy policies measured by the number of words for the top 10 and random 30 banks from 1999 to 2005. The number of data points from 1999 to 2005 for the random 30 banks are: 7, 29, 27, 26, 25, 24, and 30; for the top 10 banks are 5,8,8,5,6,5,10. the GLB privacy rule went into effect on July 1, 2001.

We observed that post-GLB policies are also more complete than pre-GLB policies. To measure completeness we looked at whether the coder was able to determine the following information for each privacy policy: what information the company collects, whether it shares with affiliates, what information it shares with affiliates, what choice it gives to consumers for affiliate sharing, and similar questions for sharing with non-affiliated third parties. We found that pre-GLB policies had a high percentage of unknowns[52], but post-GLB policies are more complete. The result for the random 30 group is presented in Figure 11. Data from the top 10 group exhibits a similar pattern; however, top-10 banks had fewer unknowns prior to GLB than the random 30 banks.[53]

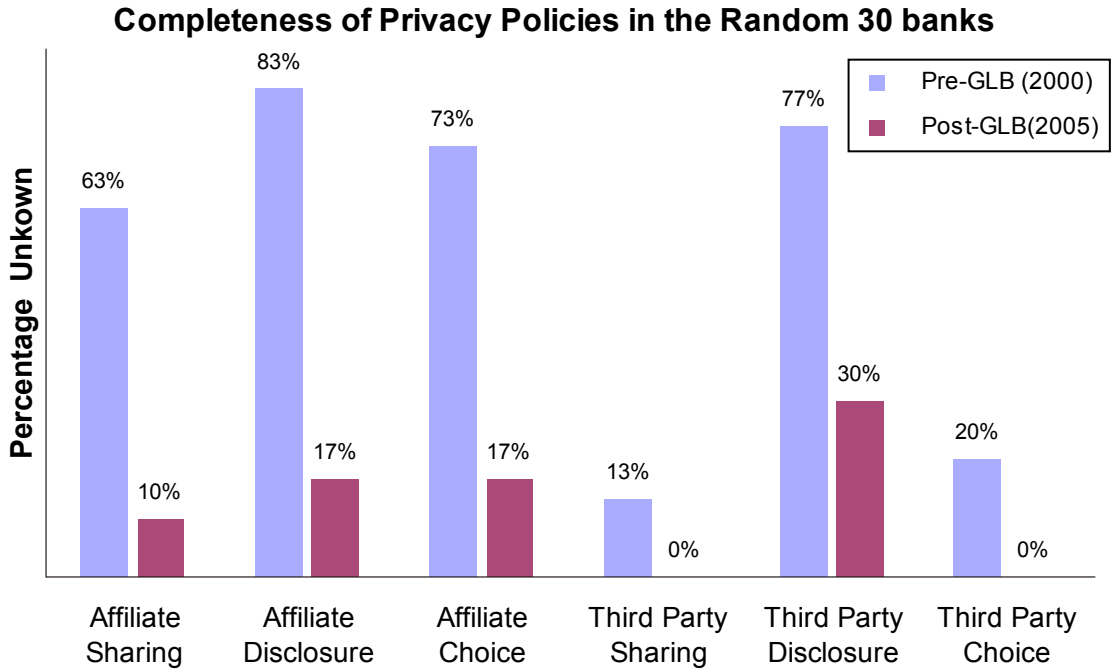

Figure 11: Percentage of privacy polices coded as "unknown" for the random 30 group. Unknown means that after reading the policy; we are unable to decide the company's practices regarding such sharing. Affiliate and Third party disclosure refers to the types of information to be disclosed to affiliates and third parties respectively.

4.2.2 Minimum GLB Compliance and Increased Information Dissemination

Top 10 banks and credit card companies surveyed appear to only minimally comply with GLB, offering few if any privacy protections that go beyond what the law requires. These companies collect large amounts of information from consumers, and share information extensively with affiliates and third parties.

Typical information that banks collect include:

- application information such as personal information, assets, income and debts;
- transaction information such as account balances, types of account, payment history, credit card usage;
- consumer report information such as creditworthiness or credit history;

- information from outside sources such as employment verification, information about credit and other relationships, verification of information such as property insurance coverage; and
- information from online interactions with the company's websites.

As shown in Figure 12, all of the top 10 banks and 90% of the top 10 credit issuers had policies in 2005 indicating that they may share *all* information they collect with the affiliates. Only 10% of the banks in the random 30 group do not share with affiliates.

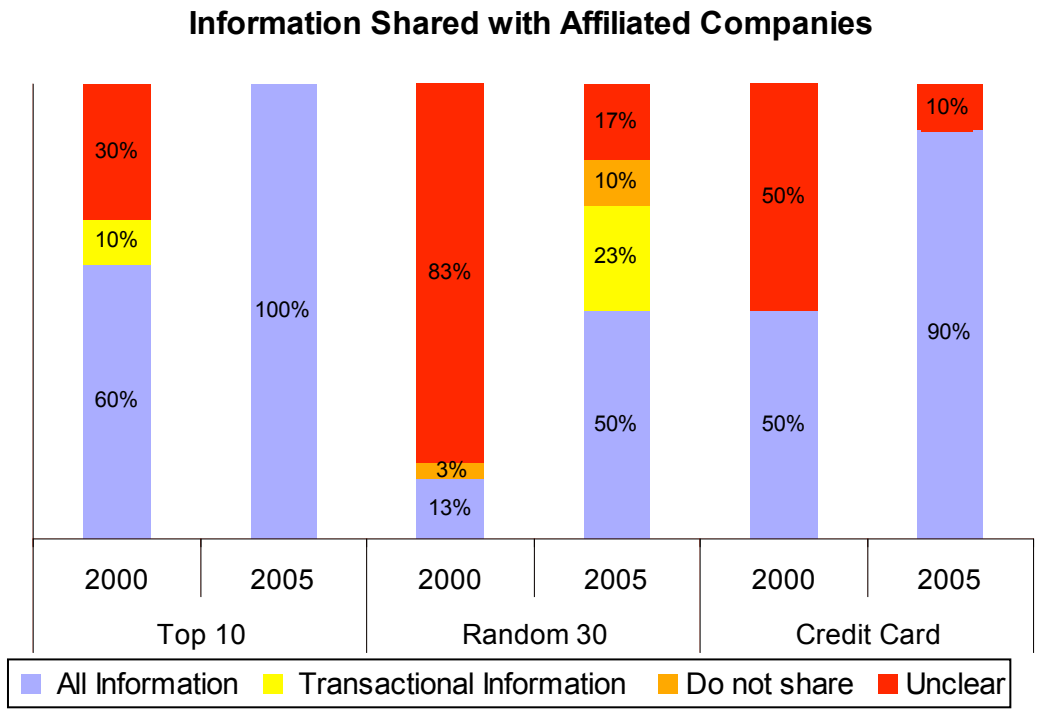

Figure 12: Affiliate sharing in 2000 and 2005 for the top 10 banks, random 30 banks and the top 10 personal credit issuers

Prior to the enactment of GLB, many bank privacy policies did not state clearly whether or not they shared information with affiliates or third parties. However, based on the high rate of affiliate sharing observed in 2005 it appears that this sharing has either increased or remained essentially unchanged. In addition, because GLB relaxes restrictions on the acquisition of affiliates, banks are acquiring an increasing number of affiliates and engaging in mergers. When banks merge, a large amount of customer information is consolidated. Since GLB places no limits on affiliate sharing, and few banks have voluntarily adopted policies that restrict their own affiliate sharing, individuals' financial data is now being shared more extensively than it was before GLB was enacted. We will address this in more detail in the policy implications section.

We also examined third party sharing (including joint marketing) before and after GLB. As shown in

Figure **13**, third party sharing also appears to have increased after GLB was enacted.

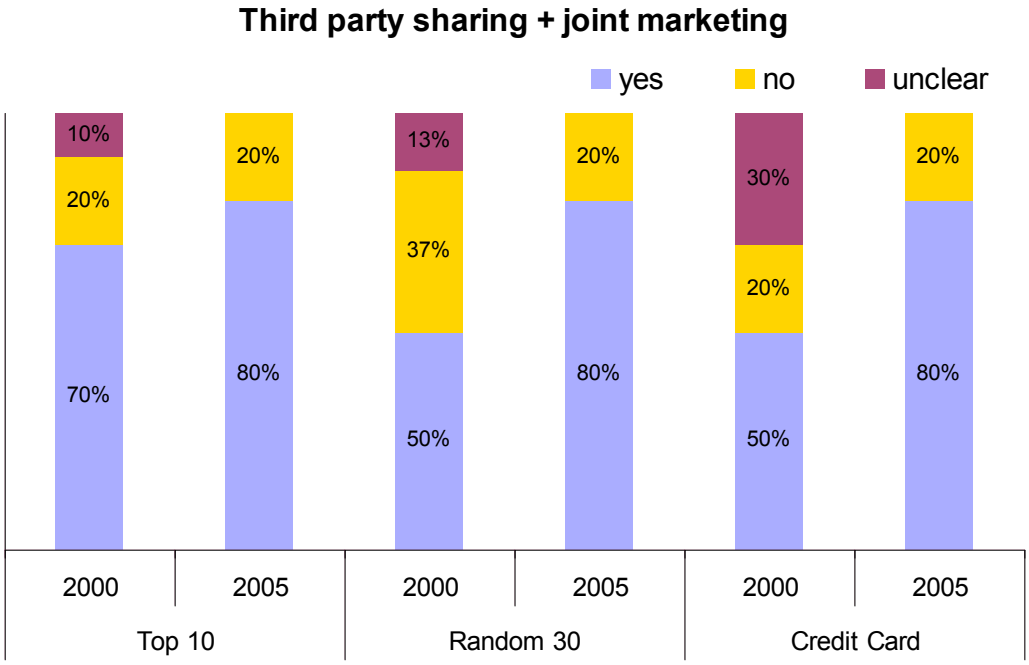

Figure 13: Third party sharing including joint marketing in 2000 and 2005 for each of the groups.

Although privacy policies often discuss third-party sharing and joint marketing agreements separately, we combine them here. We categorize joint marketing sharing as sharing with third parties, because in essence joint marketers are third parties. GLB does not require companies to offer consumers the ability to opt-out of joint marketing as long as they disclose in their privacy policies that they may engage in join marketing. However, if an opt-out is offered, then joint marketers can be treated as ordinary third parties, who may be permitted to further share the information with other companies. We find that the number of institutions sharing with third parties actual increases from 2000 to 2005.

4.3 Analysis of Readability

Did the GLB Act have an impact on the readability of financial privacy notices? Some regulations may require disclosures that are so complicated that it is almost impossible to comply without producing notices that are difficult to read. Other regulations may be accompanied by guidance that results in more readable notices. A number of studies have demonstrated that financial privacy notices are difficult to read and understand. Are the current readability problems with GLB notices likely caused by the regulation, or have the notices actually improved (or remained unchanged) as a result of the regulation?

4.3.1 Gradual but Insignificant Improvement

The readability of financial privacy policies continually improved during the time period examined, with the largest changes observed between 1999 and 2002. In 1999, policies had an average Kincaid readability score of 14.5. This dropped to 13.0 in 2002. Figure 14 shows this improvement for the top10 and random 30 groups of banks.

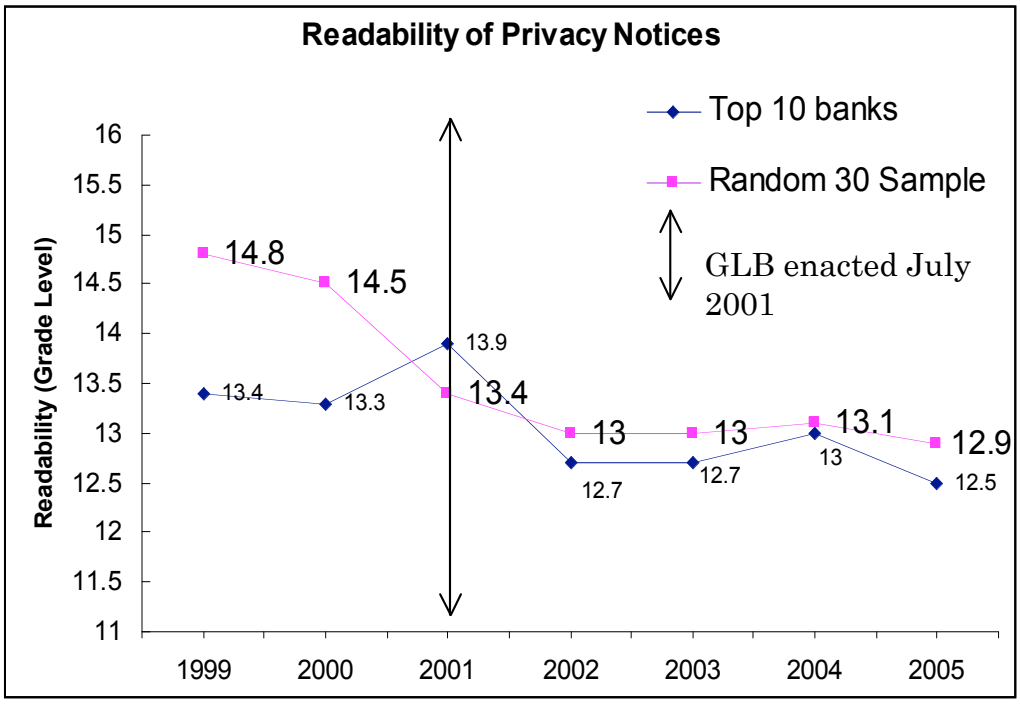

Figure 14: Mean readability of privacy policies measured by the Kincaid score for the top 10 and random 30 banks group from 1999 to 2005. The number of data points for the random 30 banks from 1999 to 2005 is: 7, 29, 27, 26, 25, 24, and 30; for the top 10 banks are 5,8,8,5,6,5,10. GLB went into effect on July 1, 2001. The Kincaid score corresponds to the equivalent years of education needed to understand the text.

The average Kincaid score after 2002 leveled off at about 13.0, which roughly is equivalent to the U.S college freshman reading level.[54] In light of the fact that in 2005, 48% of the population over 15 has a high school or less education,[55] and comparing the common readability scores for other materials, the readability improvement we observed probably has not had much real impact. In addition, it is important to keep in mind that readability scores are calculated based on the length of words, sentences, and paragraphs only. They do not take into account use of jargon or unfamiliar words, vague language, complicated sentence structure, or references to laws with which most people are unlikely to be familiar.

For example, although the following two statements have similar readability scores, the second makes a much more direct statement than the first:

Bank A: "We do not share information about you with third parties outside of … , except as permitted by law."

Bank B: "We may share all the information we collect with our affiliates, and we may also share your information with companies we have joint marketing agreements with."

Arguably Bank A's statement is likely to mislead most consumers into thinking that Bank A engages in little sharing of personal information. In fact, Bank A and Bank B have virtually the same data sharing practices. The reason for this misunderstanding is that consumers may not know that the law actually contains only very limited restrictions on data sharing.

4.3.2 Change in Model Policy

From reading the privacy policies, we observed that before GLB, many financial institutions posted privacy policies that modeled the American Bankers Association's privacy principle.[56] The principle, announced in 1997, was an industry wide initiative to promote privacy protection.[57] Pre-GLB, many banks—especially those in the random 30 group—based their privacy policies closely on the principle, making few modifications. The FTC, in its final rule for GLB privacy in 2000, included GLB compliant privacy policy examples. Post-GLB, most banks use the FTC's template with little modification.

We verified our observation using a textual similarity measure. Our measure gauges the similarity between the two policies by the number of overlapped words and the frequency of their usage. If two documents use the same set of vocabularies and the frequency of their usage is similar, then our measure would be close to 1.

Our analysis of shows that there is a significant change in the content of the privacy policies from 2000 to 2001 as they use a different set of vocabularies. This is indicated by the fact that the similarity score for policies from 2000 to 2001 is much lower than in the other years we examined. The results are shown in Figure 15.

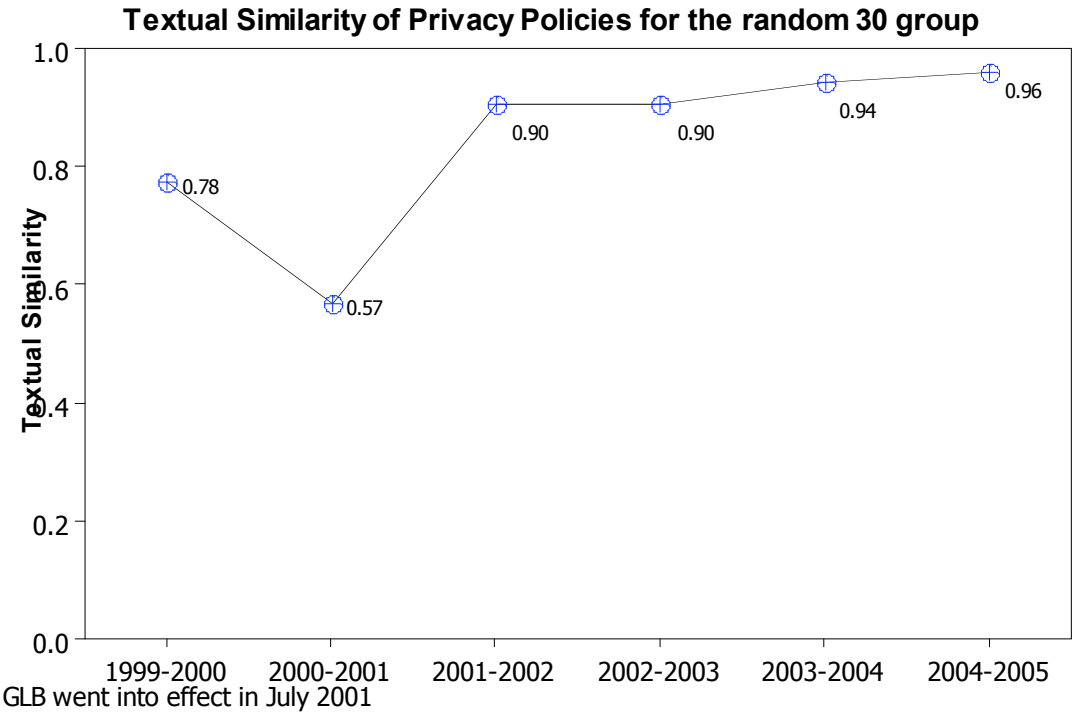

Figure 15: Mean textual similarity of privacy policies for the random 30 group. Policies from the same institution in adjacent years are compared (for example 1999 with 2000, 2000 with 2001)

5 Platform for Privacy Preferences

The Platform for Privacy Preferences (P3P) is a standard format for privacy policies, and is formally recommended by the World Wide Web Consortium (W3C).[58] P3P provides an XML hierarchy in which privacy policies can be expressed. By using a standard XML format, people can more readily understand exactly which privacy protections a given website offers. Moreover, P3P user agents can read P3P policies automatically, and can report when privacy policies meet a user's criteria. This eliminates the need for people to read each privacy policy to determine if a website has adequate privacy protections.

5.1 Overview of P3P History

P3P was inspired, in part, by the Platform for Internet Content Selection (PICS) system. PICS is a W3C standard designed to enable parents to limit their children's Internet browsing.[59] For example, parents can select settings in Internet Explorer to reject all websites that reported pornographic content. PICS gained additional notoriety when it became central to the argument against the Communications Decency Act of 1996 (CDA.) The United States Congress passed the CDA to protect minors from objectionable Internet content. The courts overturned the CDA because it failed the least restrictive means test. Rather than government censorship of the entire Internet, technological solutions like PICS could tailor filters specifically for children.[60]

The Center for Democracy and Technology (CDT) worked with the Federal Trade Commission (FTC) to promote the idea of a tool like PICS that would enable customers to filter out websites that had unacceptable privacy protections.[61] P3P was created with the support of corporate partners including AT&T and Microsoft. P3P was developed by W3C working groups over a five year period, and accepted by the W3C in 2002.[62]

5.2 P3P Adoption Rates

This section is excerpted with minor modifications from *An Analysis of P3P-Enabled Websites among Top-20 Search Results*.[63]

Users frequently use search engines to locate information on the Internet. Search engines have taken on the role of "gatekeepers of the web".[64] 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.[65]

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.

We also obtained a list of the 30,000 most clicked on domains from AOL search results collected during October of 2005. This list included the number of clicks made to each domain during that period. We checked each of these domains for P3P policies. Of the 30,000 domains, 2,564 unique domains (8.54%) had P3P policies. However, examining the number of clicks to these sites, we found that these 2,564 domains accounted for 16.67% of the total traffic.

We collected the first twenty hits from every search term during the summer of 2005. We conducted Privacy Finder searches with all of the terms in the AOL and Froogle data sets using both the Google and Yahoo! APIs. We also collected the first twenty hits obtained using AOL's search engine for the terms in the AOL data set. For every search term returned, we checked for the existence of a P3P policy. For the sites that did have policies, we then evaluated them against five APPEL rule sets. 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.

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.

5.2.1 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) went to URLs that had P3P policies available (10.14%). However, not all of these policies are unique; many of the hits come from multiple different pages on a single domain. In many cases, multiple domain names use the same policy, often because they are owned by the same company. So of the 113,880 P3P-enabled search hits, we found only 3,846 unique policies.

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

Overall, there are a relatively small number of sites that 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 16.

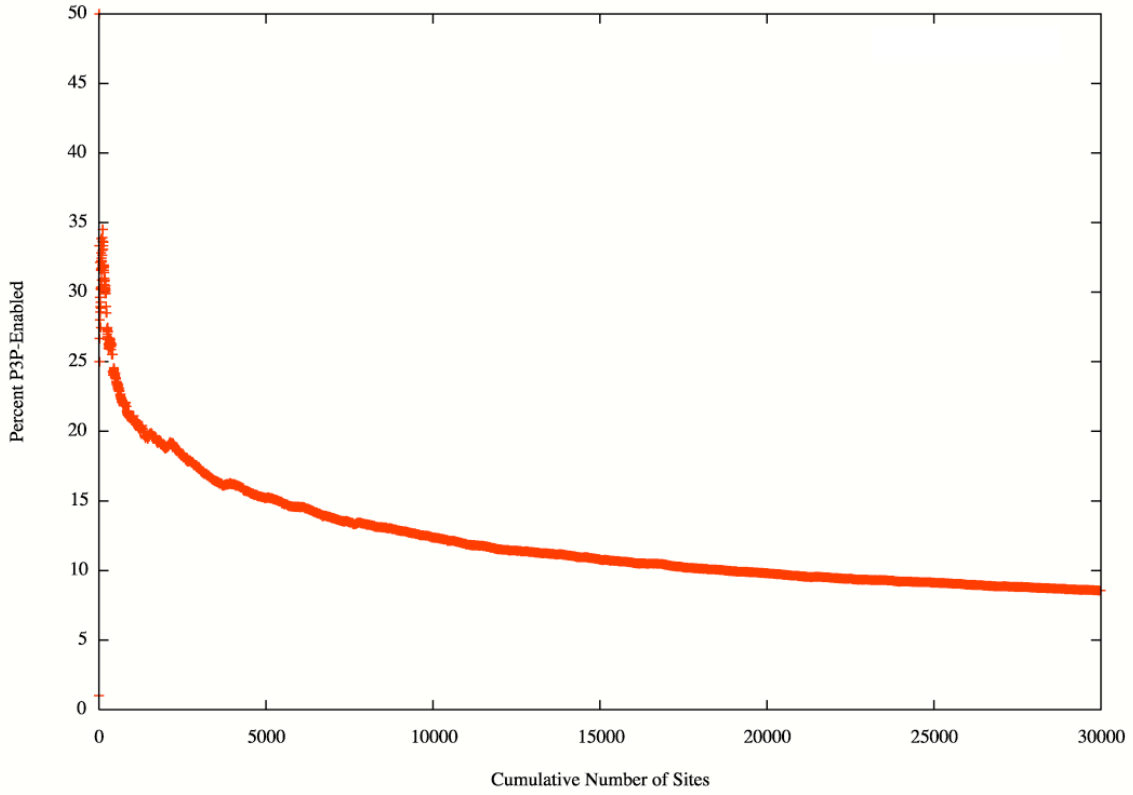

Figure 16: Graph of the number of P3P-enabled sites across the most popular websites. For instance, 15% of the top 5,000 most popular sites are P3P-enabled.

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 17.

Overall, there are a relatively small number of sites that get returned by the search engines frequently. Specifically, the top twenty most popular P3P-enabled domains account for over 50% of the total number of P3P-enabled hits discovered. The rate at which pages are returned seems to follow a Zipf-like distribution (the frequency trend follows a power law), as shown below in Figure 17. This distribution is very similar to the one depicted in Figure 16, where we examined the 30,000 most popular domains.

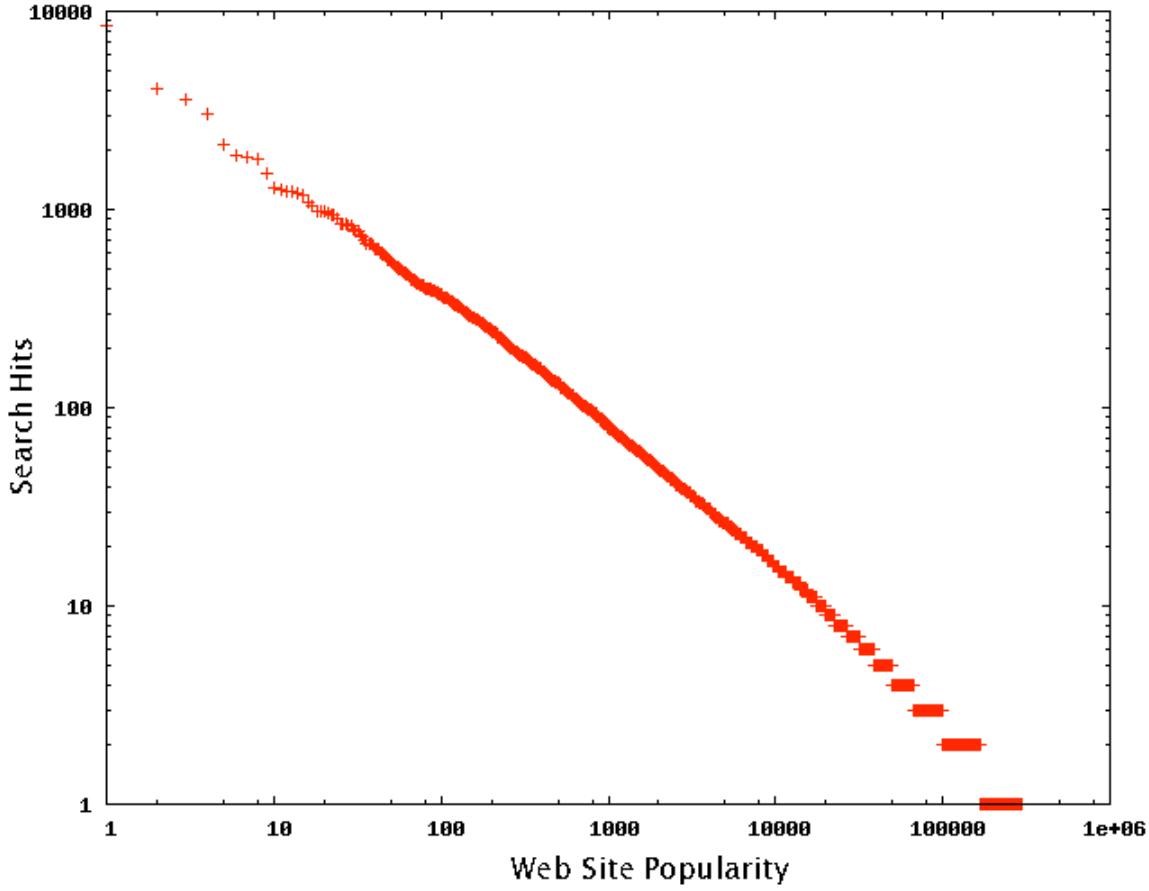

Figure 17: Plot of website frequency following a power law. This data reflects the search results yielded from the 20,000 AOL search terms.

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

5.2.2 P3P-Enabled Websites in Search Results

As part of our examination of P3P adoption, we investigated the frequency with which popular search engines return P3P-enabled search results.

Google functions by examining the number of links to a particular page, the text on those links, and the number of links to those linked pages.[66] 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.

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

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 6. 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 ($p < 0.0005$).

Of all the typical search terms, only 638 of them yielded no results across all three search APIs. This amounts to roughly three percent. We also found that there are a small number of P3P policies that are likely to appear in a large number of search queries. Of these, Yahoo!'s P3P policy is the most prevalent. Overall, there were 31,905 search hits that used this policy, corresponding to 23,335 URLs found on 4,015 different host names. This is because in addition to running a search engine, Yahoo! also offers web hosting services. Thus, when a hosting customer creates a site, they will automatically be using Yahoo!'s P3P policy.[67]

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 7 shows the top ten P3P policies using both data sets.

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

5.2.3 Longitudinal Trends

In the summer of 2003, Byers, et al. conducted the first automated study of P3P adoption.[68] 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.[69] 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.[70] Another list used was the comScore Media Metrix Key Measures, another top 500 list that also included third parties such as advertisers. Another list contained the top 500 domains from the Alexa Traffic Ranking as of February 2003.

The last four lists were created by the researchers after crawling various sites. Froogle was used to create a list of 1,017 commerce-related sites.[71] Yahooligans!, a web index run by Yahoo! and geared towards children ages 7-12, was used to create a list containing 900 sites. FirstGov was crawled to create a list of 344 U.S. government websites. Finally, Google News was crawled to create a list of 2,429 news-reporting sites. In total, 5,856 unique sites were examined, 588 of which were P3P-enabled. In addition to comparing our search engine data with this data, we also re-examined the lists of sites used in this previous study. Our findings can be seen in Table 8.

	# in list	Sites reached in 2003	P3P-enabled in 2003	Sites reached in 2006	P3P-enabled in 2006	% 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 8: Revisiting the 2003 study on P3P adoption

Of the 5,856 unique sites examined, 5,739 were accessible in 2003, and 5,414 were accessible when we repeated this study in February of 2006. The results here show that overall there was an increase in total P3P adoption over the 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 which mandates government agencies post machine-readable privacy policies on their websites.[72] Additional increases can be seen with regard to news-related sites as well as websites targeted at children.

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 Geographic Distribution of P3P Policies

P3P enjoys international deployment, but is still predominately concentrated in the United States. We used our cache of 9,408 P3P policies to explore international P3P adoption trends. We identified 437 non-US P3P-enabled sites.

We calculated the number of P3P-enabled websites from a given country represented in our cache by examining the ccTLDs (country code top level domains — for example, .uk is the ccTLD for the United Kingdom). In addition to .us, we assumed all .com, .mil, .org, .net, .gov, .edu, .info, and .biz TLDs are in the United States. However, international companies do own a subset of those domains, particularly since .com is seen as the most desirable top level domain. Consequently our results may slightly undercount non-US P3P-endabled sites and should be seen as a minimum for P3P diffusion.

In the other direction, we may incorrectly classify some non-US sites due to US entities who purchase desirable ccTLDs like .tv, which the US company Idealab bought from the country of Tuvalu[73]. Similarly .cc, which is the country code for the 574 inhabitants of the Cocos Islands,[74] is marketed by the Seattle-based eNIC corporation.[75] Considering how few sites we have with .tv and .cc ccTLDs, on balance it is likely we that overall we have undercounted non-US P3P-enabled sites.

Of the 437 non-US P3P-enabled sites, the lion's share are in the United Kingdom. Other English-speaking countries, Australia and Canada, are also on the top five list. The second and fifth highest rates of P3P adoption are in two nations where English is not the primary language: Japan and Germany.

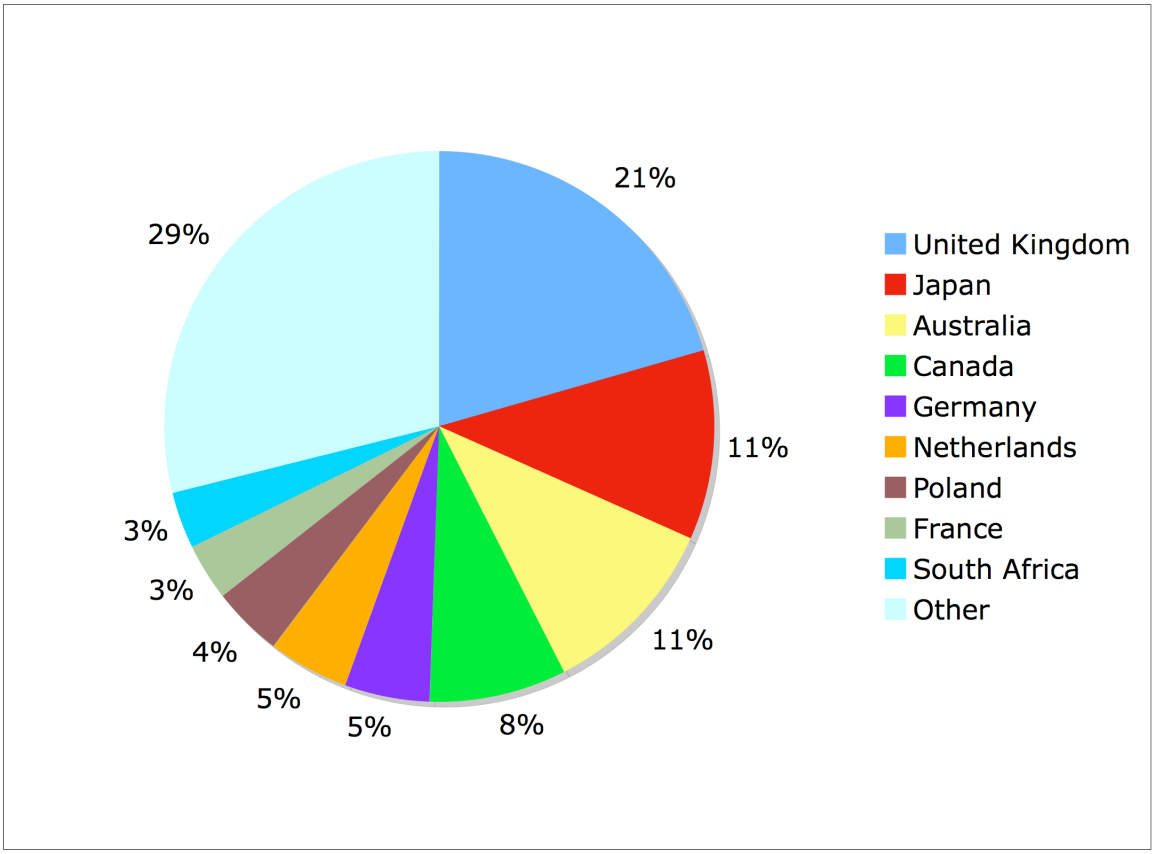

Figure 18: Top P3P adoption rates outside the US. Nations listed in order of number of P3P-enabled websites

In all, forty eight different countries in our sample had at least one P3P-enabled website. The breadth of P3P's reach is impressive. Forty-five percent of Non-US countries with P3P-enabled sites are member states of the European Union. EU member nations are subject to the provisions of Directive 95/46/EC of the European Parliament, which provides protections for personal data privacy.[76]

One potential avenue of future research is to survey the owners of non-US P3P-enabed websites to understand how they learned of P3P and why they adopted it. The diversity of nations with P3P sites shows P3P has a broad appeal. Understanding P3P's diffusion may help identify ways to promote greater depth of P3P adoption.

TLD	Country	# of Sites	% of Non-US P3P sites	EU Member
.uk	United Kingdom	91	20.8%	✓
.jp	Japan	49	11.2%	
.au	Australia	48	11.0%	
.ca	Canada	35	8.0%	
.de	Germany	22	5.0%	✓
.nl	Netherlands	22	5.0%	✓
.pl	Poland	18	4.1%	✓
.fr	France	15	3.4%	✓
.za	South Africa	14	3.2%	
.br	Brazil	11	2.5%	
.tv	Tuvalu	11	2.5%	
.es	Spain	9	2.1%	✓
.mx	Mozambique	9	2.1%	
.kr	Republic of Korea	7	1.6%	
.it	Italy	6	1.4%	✓
.nz	New Zealand	6	1.4%	
.ar	Argentina	4	0.9%	
.cl	Chile	4	0.9%	
.cn	China	4	0.9%	
.dk	Denmark	4	0.9%	✓
.hk	Hong Kong	4	0.9%	
.ru	Russian Federation	4	0.9%	
.in	India	3	0.7%	
.ac	Ascension Island	2	0.5%	
.at	Austria	2	0.5%	✓
.be	Belgium	2	0.5%	✓
.gs	South Georgia	2	0.5%	
.il	Israel	2	0.5%	
.no	Norway	2	0.5%	
.ro	Romania	2	0.5%	
.se	Sweden	2	0.5%	✓
.sg	Singapore	2	0.5%	
.to	Tonga	2	0.5%	
.tw	Taiwan	2	0.5%	
.ve	Venezuela	2	0.5%	
.ae	United Arab Emirates	1	0.2%	
.az	Azerbaijan	1	0.2%	
.cc	Cocos (Keeling) Islands	1	0.2%	
.ch	Switzerland	1	0.2%	
.cz	Czech Republic	1	0.2%	✓
.gr	Greece	1	0.2%	✓
.ke	Kenya	1	0.2%	
.md	Republic of Moldova	1	0.2%	
.my	Malaysia	1	0.2%	
.pe	Peru	1	0.2%	
.pr	Puerto Rico	1	0.2%	
.pt	Portugal	1	0.2%	✓
.ws	Samoa	1	0.2%	
Total Non-US TLDs		**437**	**100%**	**45%**

Table 9: International P3P deployment

5.3.1 Differences in Data Collected By European Union and Other Nations

The high-level picture is unambiguous: European Union websites have more privacy protective practices than non-EU websites outside the US. Unlike the Gramm-Leach-Bliley Act, the EU privacy directive may be responsible for enhanced privacy protections. Right now we have only established correlation. It would be interesting to look at changes in EU policies over time as we did with the financial industry to see if the EU websites were already more privacy protecting prior to the privacy directive, or if there is a causal relationship.

5.3.2 Types of Data Collected

Nearly all sites collect both computer and navigation data. While EU sites collect such data less frequently, there is not much difference. However, in all other significant categories EU sites collect fewer types of data. Location and Other are the only two exceptions, but neither EU nor non-EU sites collect enough data in those categories for the differences to be significant. In many data categories EU sites are only half as likely to collect data as their non-EU counterparts.

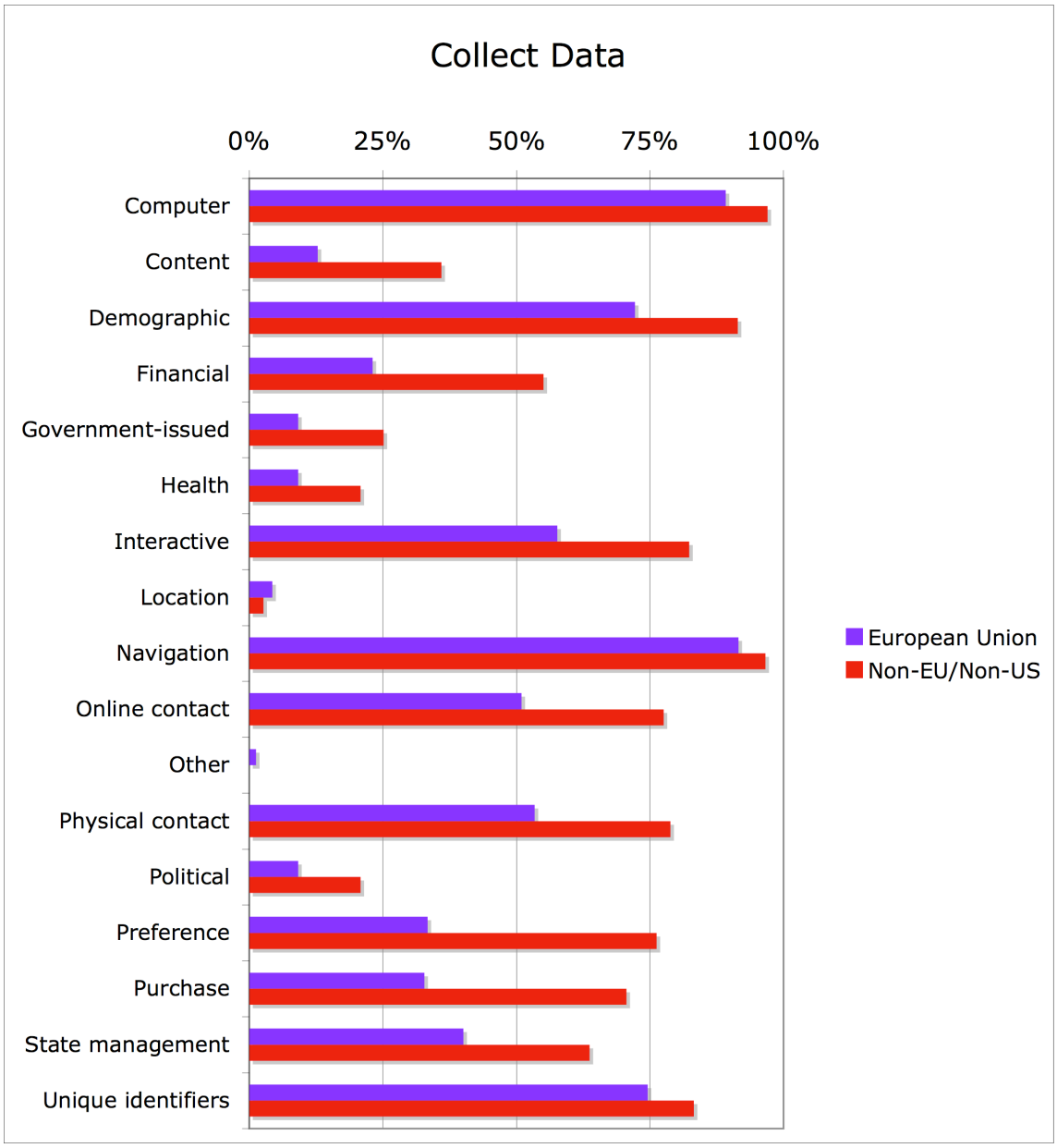

Figure 19: Data Collection of Non-US Websites by European Union Membership

5.3.3 Data Use

Support of the current activity, system administration, and research and development remain the three most prevalent uses for data. EU sites are less likely to use data for marketing, telemarketing, and customizing the web browsing experience. They are generally more privacy protective than non-EU sites, although the differences are not as dramatic as the differences we saw in data collection practices.

Figure 20: Data Use of Non-US Websites by European Union Membership

5.3.4 Data Recipients and Sharing

This graph may be the most striking in the series. For all categories, EU sites are less likely to share or sell information about their visitors.

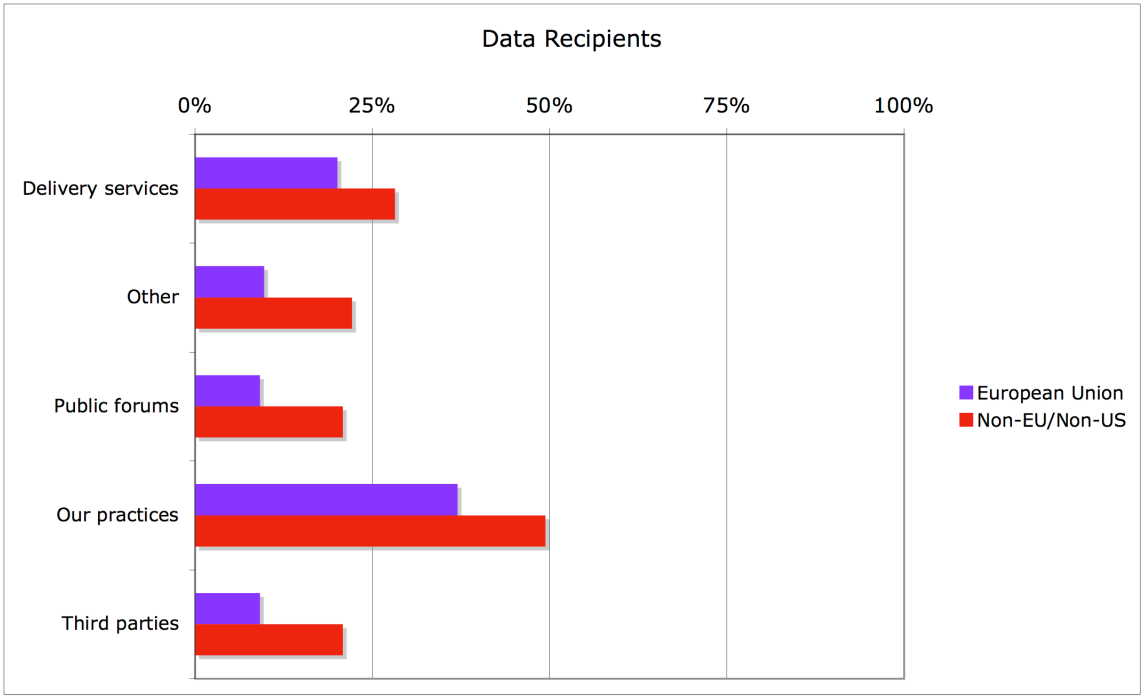

Figure 21: Data Recipients of Non-US Websites by European Union Membership

5.3.5 Access Provisions

European Union sites provide superior access to data, so customers can see their own data to check for errors. EU sites are more likely to not collect any identifiable information, and more likely to provide access to all data. Consequently fewer EU sites provide access to contact information, because they either do not collect it or provide more than that level of access.

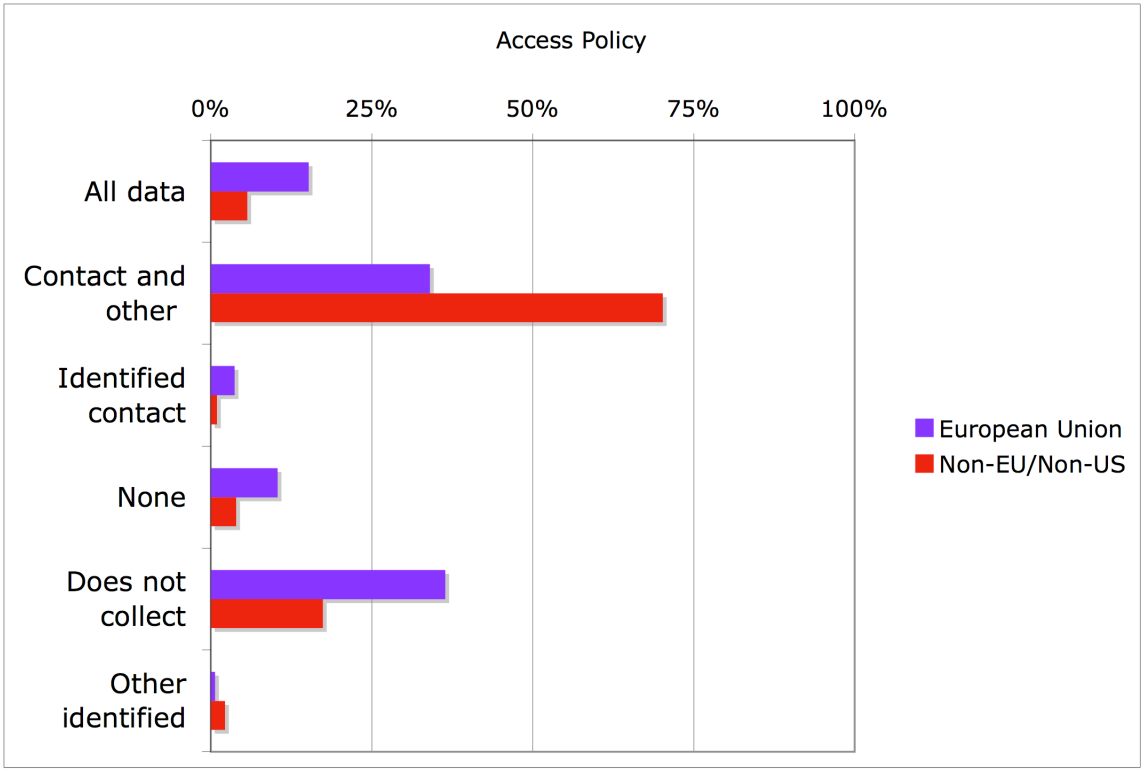

Figure 22: Access Policies of Non-US Websites by European Union Membership

5.3.6 Dispute Resolution Options and Remedies

Across the board, customer service is the most popular suggestion for resolving disputes. Independent organizations such as the TRUSTe[77] and Better Business Bureau[78] seals are more popular with non-EU sites, and appear to supplant customer service for some sites. Courts remain the least popular mechanism, and applicable laws are rarely mentioned.

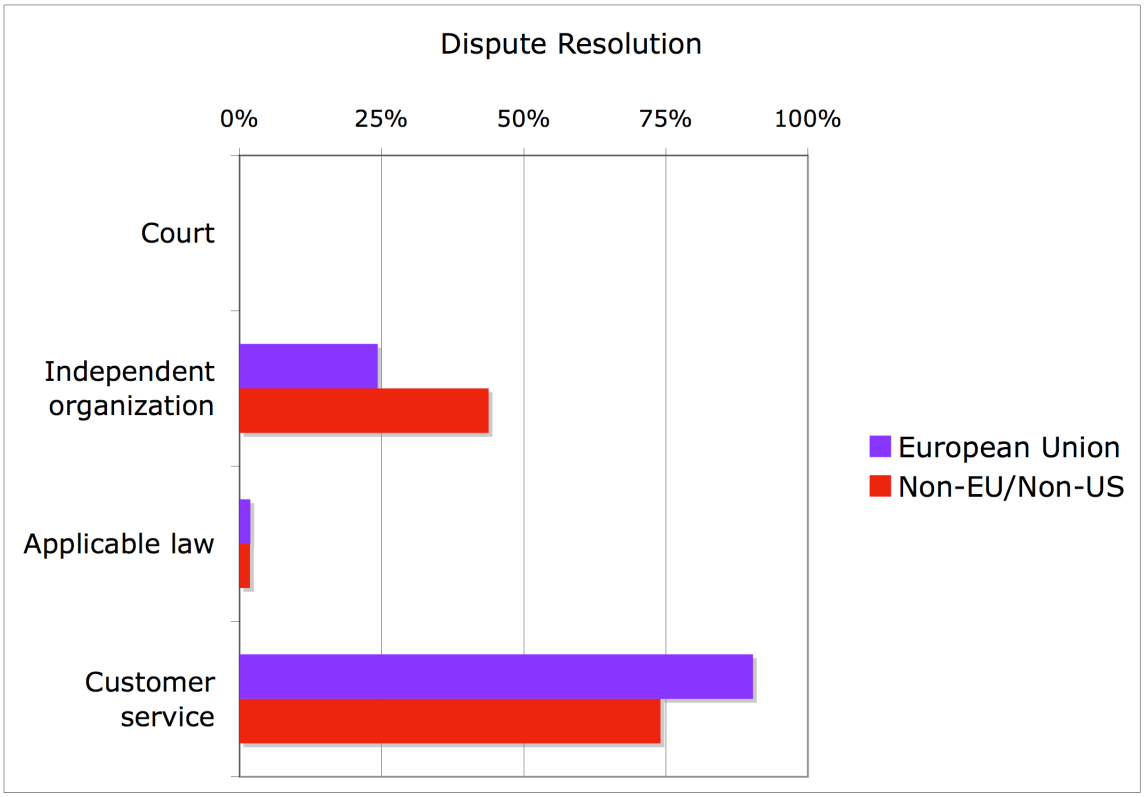

Figure 23: Dispute Resolution of Non-US Websites by European Union Membership

5.3.7 Data Retention Policies

As in the United States, websites retain data for a very long time and many do so indefinitely. Most sites retain data not out of legal obligation, but for business reasons. In this case the EU sites are less privacy protective than the non-EU sites, but the two are fairly similar.

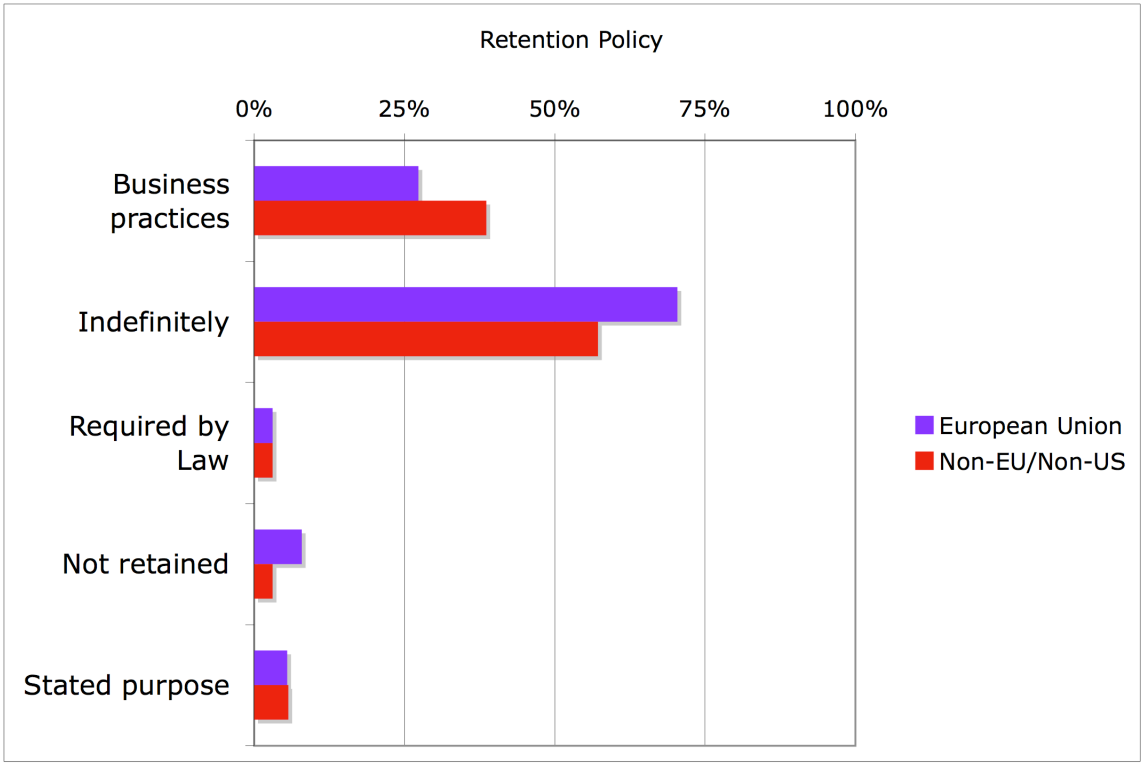

Figure 24: Data Retention Policy by European Union Membership

We used Yahoo! directory categories[79] to analyze P3P adoption by market segment. We found dramatic differences in adoption rates, from nearly 20% for websites targeted to selling products to consumers, to only 0.5% adoption for education.

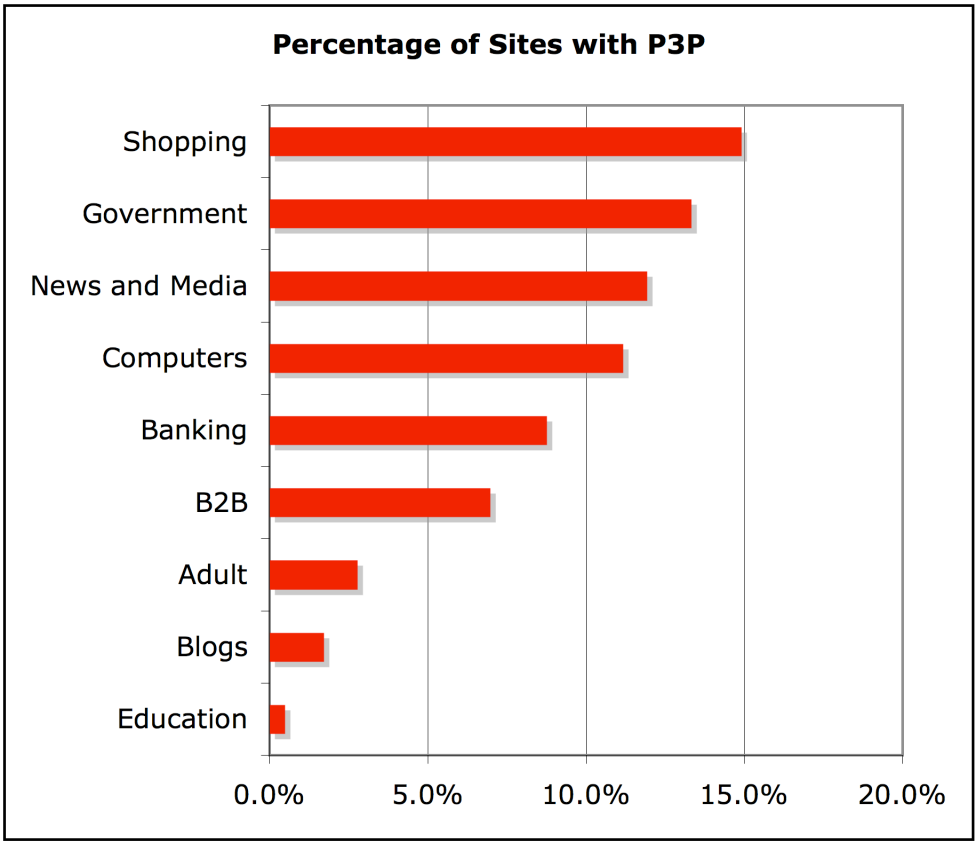

Figure 25: Percentage of Sites with P3P by Yahoo! Directory Category

Yahoo! directory categories are mutually exclusive; even if a website spans two categories it will only be counted in one. We automated mapping URLs from the AOL dataset to the websites in Yahoo! categories as of October, 2006. A few categories may benefit from additional explanation.

Computers: The "computers" category encompasses non-retailers. These are sites with information on computers, technology, and so forth.

B2B: B2B stands for business-to-business. This refers to sales to corporations, rather than end customers. For example, if General Motors purchases paperclips in bulk, that is a B2B transaction. Many large B2B transactions involve corporate purchasing agents and may be limited to negotiations with a handful of trusted suppliers.

Education: In addition to the sites in the Yahoo! category for education, we also coded all sites with a .edu TLD (top level domain) as well as sites that were in the "university" or "college" categories as part of the education category.

We notice a few interesting trends:

- P3P adoption is not simply a function of financial transactions. While both Shopping and B2B sites are engaged in selling products, Shopping sites have twice the adoption rate as B2B sites. Banking sites fall in between.
- P3P adoption may, in part, be driven by the level of computer-savviness of the organization. Even though sites in the Computer category do not need to worry about losing sales by not offering a P3P policy, they still have a high adoption rate.
- Education lags in P3P adoption, even behind adult sites and blogs.

In total, we analyzed 16,919 sites in these nine categories, representing 6% of our AOL sample, as shown in Table 10. There are 143,080 sites that fall into categories other than these nine.

Category	Number of Sites	Number with P3P	Percentage with P3P
Shopping	2787	415	14.9%
Government	3050	406	13.3%
News and Media	1351	161	11.9%
Computers	278	31	11.2%
Banking	366	32	8.7%
B2B	1049	73	7.0%
Adult	722	20	2.8%
Blogs	646	11	1.7%
Education	6670	32	0.5%
	Total: 16,919	**Total: 1,181**	**Average: 8%**

Table 10: P3P-enabled sites across categories found using the Yahoo! Directory.

5.3.8 Types of Data Collected

Nearly all sites collect both computer and navigation data. The fact that blogs appear to collect computer data less frequently may be an indication that some blogs are hosted in such a way that the data lies with the ISP and not the blog owner, rather than a desire to protect visitors' privacy. More likely, though, it may reflect errors in coding P3P. We would expect to see all sites collect data in the computer category.

Shopping sites collect a greater breadth of data. Perhaps the most surprising is how many shopping sites collect political information. This is likely because the shopping sites use the Yahoo! privacy policy, which indicates political information is collected. It is likely that fewer shopping sites actually collect political information than report they do so.

Nearly a quarter of banks 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 a server. Such information could be useful for fraud detection, but we have no knowledge that this is happening. This might be an interesting area for further study. Once again, government sites collect the fewest forms of data.

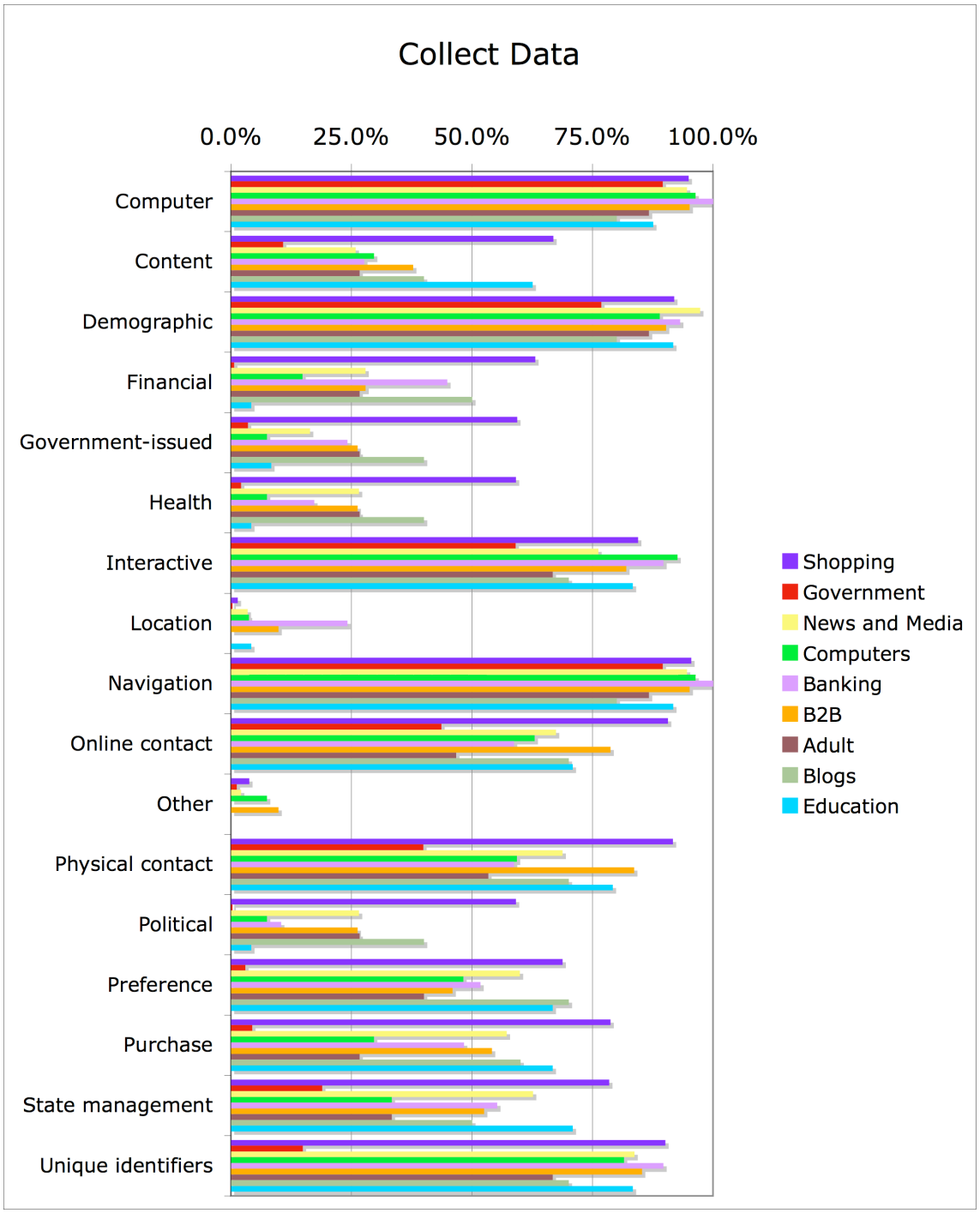

Figure 26: Data Collection by Type of Website

5.3.9 Data Use

Support of the current activity, system administration, and research and development remain the three most prevalent uses for data. Shopping sites use data in just about every way possible, where government sites are more restrictive. We are surprised to see adult sites report they store data for "historical preservation," and suspect they may misunderstand that P3P category. We would expect that to be highest for government. We are also surprised to see high rates for telemarketing. This differs from the lower bounds on telemarketing we saw for the Popular and Random sites, and suggests telemarketing occurs but is unmentioned in human readable policies.

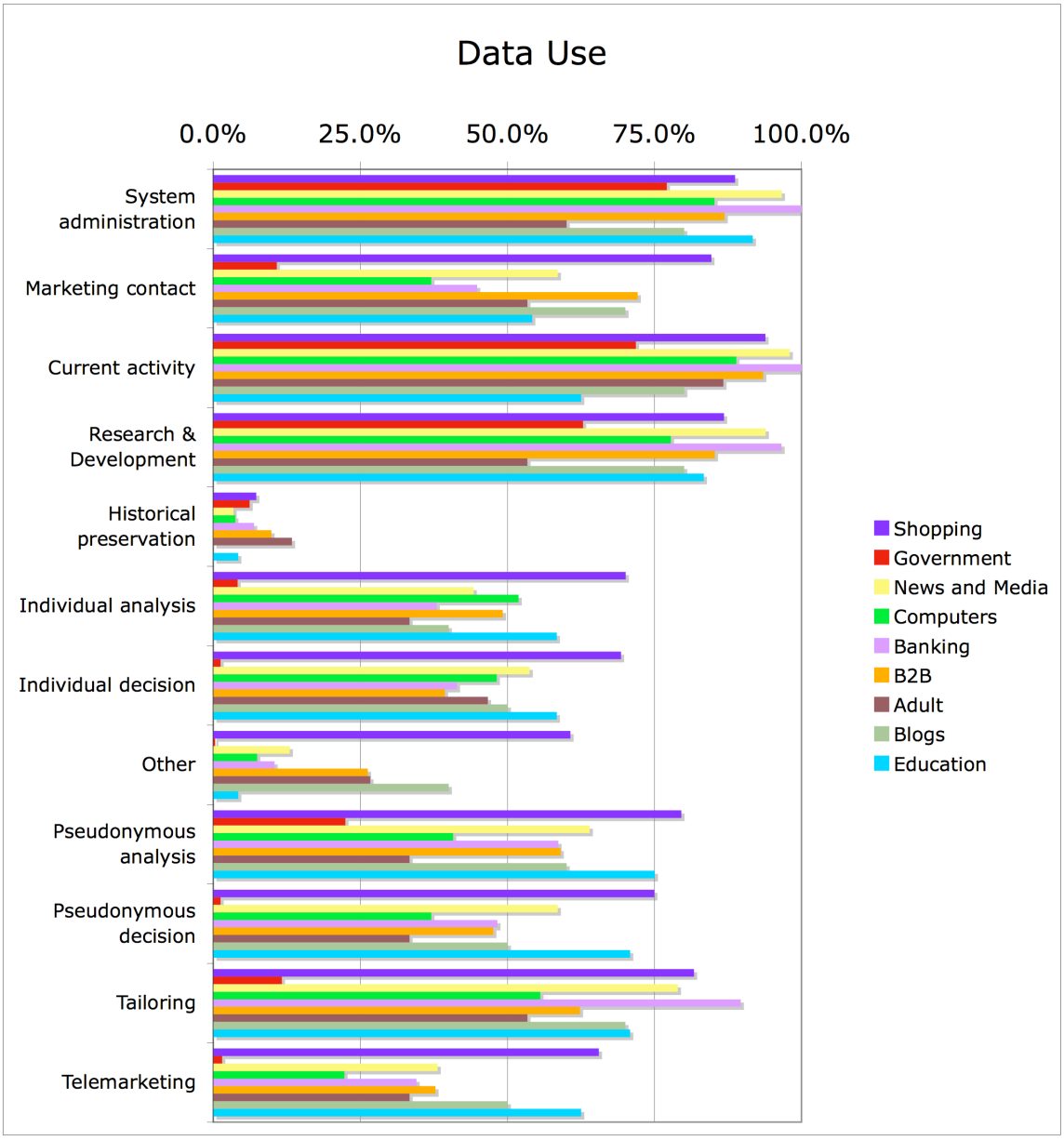

Figure 27: Data Use by Type of Website

5.3.10 Data Recipients and Sharing

Once again shopping sites share data more widely than any other sector. Government does not share data frequently. It is not immediately obvious why over a quarter of blog sites share information with delivery services. As with shopping sites using the Yahoo! P3P policy, this may be due to blogs using a common policy regardless of whether the particulars apply. On the other hand, we expected the public forums category to be quite high, given the very nature of blogs.

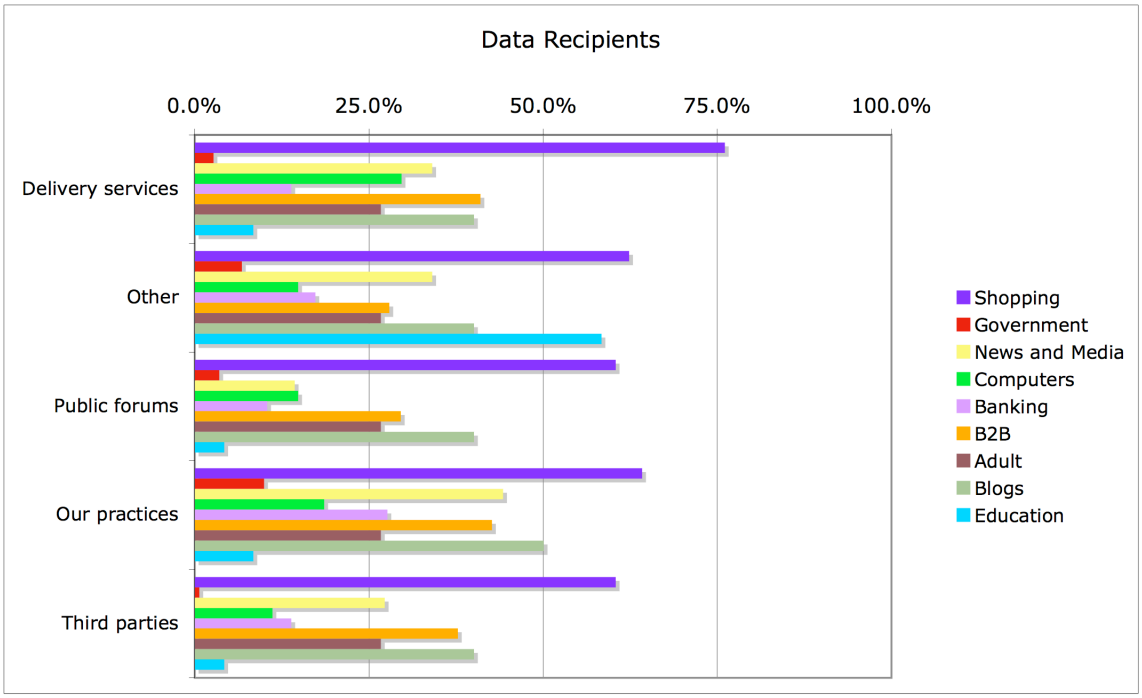

Figure 28: Data Recipients by Type of Website

5.3.11 Access Provisions

While shopping sites collect the most information, they also provide the greatest level of access to basic contact information. However, they do not provide access to all data, which includes information about how and why decisions are made about customers (recall the high rates for individual analysis and individual decisions.) Government and education sites are least likely to grant access to data, but also the least likely to collect data in the first place.

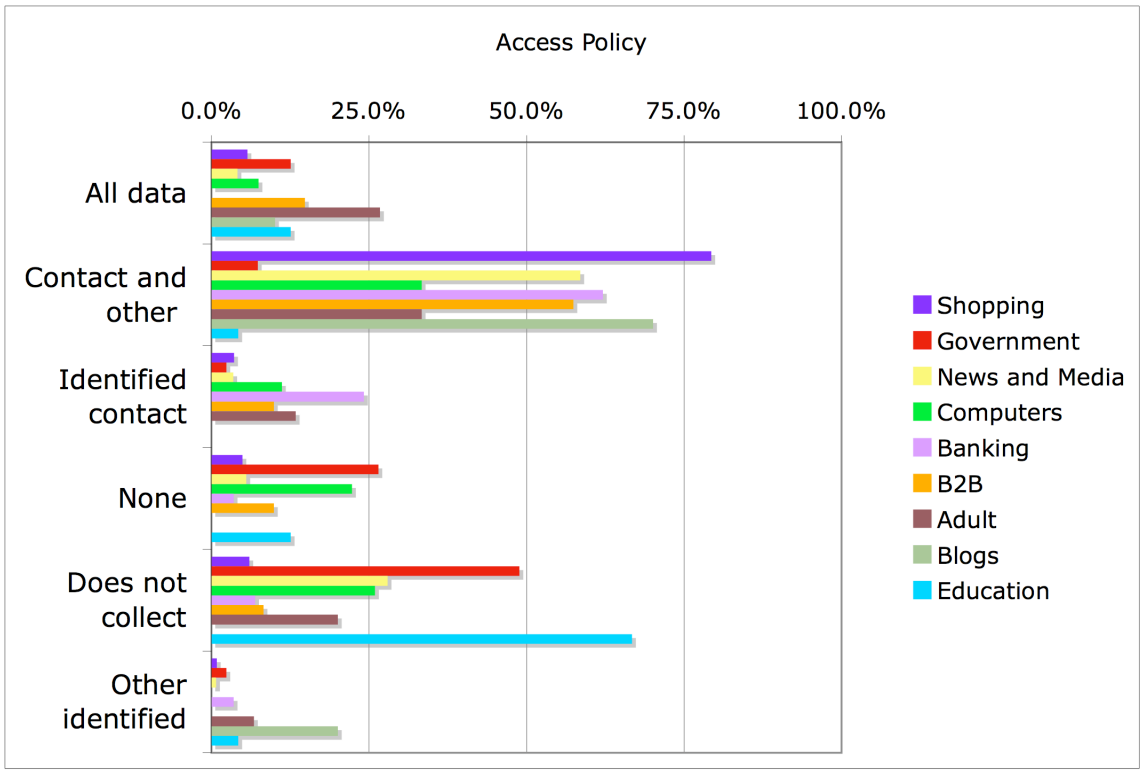

Figure 29: Access Policies by Type of Website

5.3.12 Dispute Resolution Options and Remedies

Across the board, customer service is the most popular suggestion for resolving disputes. Independent organizations such as the TRUSTe[80] and Better Business Bureau[81] seals are most popular with shopping sites. It is surprising that over half of the sites in the blog category also report they work with independent organizations. Government is notably behind in adopting mechanisms from independent organizations, perhaps relying on their own innate credibility instead. Courts remain the least popular mechanism, and government sites are the most likely to inform visitors of applicable laws.

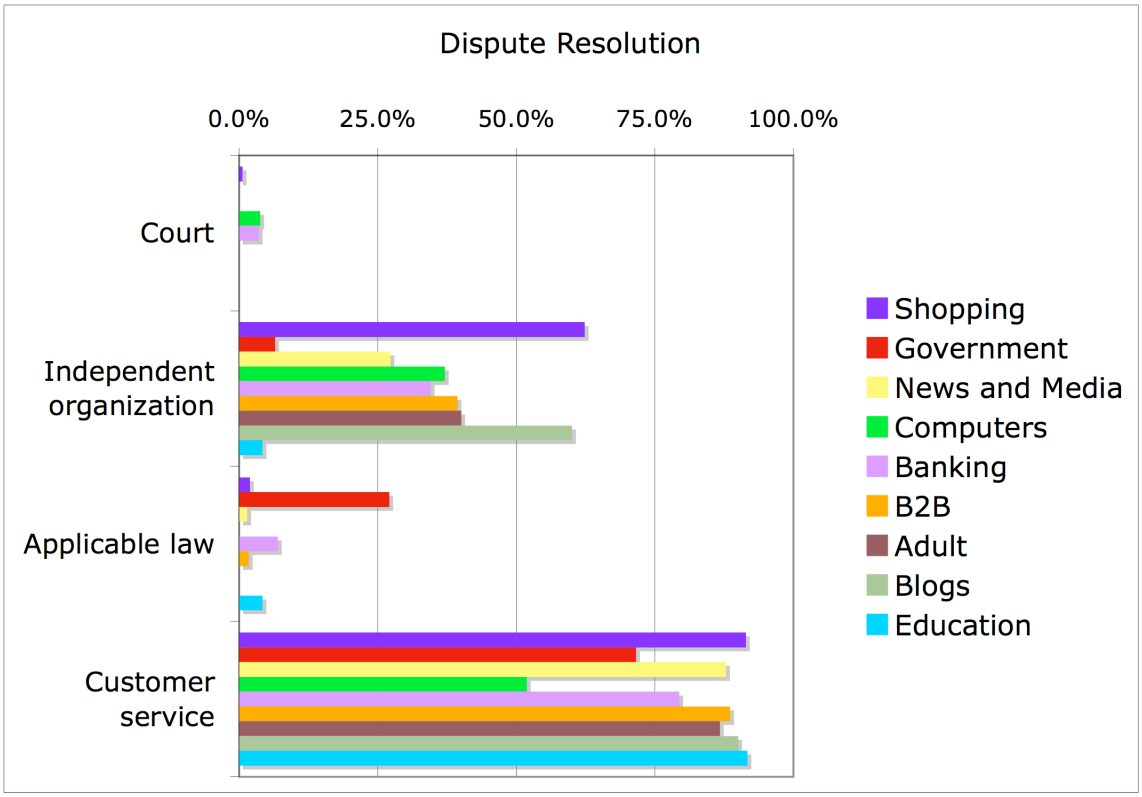

Figure 30: Dispute Resolution by Type of Website

5.3.13 Data Retention Policies

Companies retain data for a very long time, with the vast majority of educational sites and shopping sites retaining data indefinitely. Most sites retain data not out of legal obligation, but for business reasons.

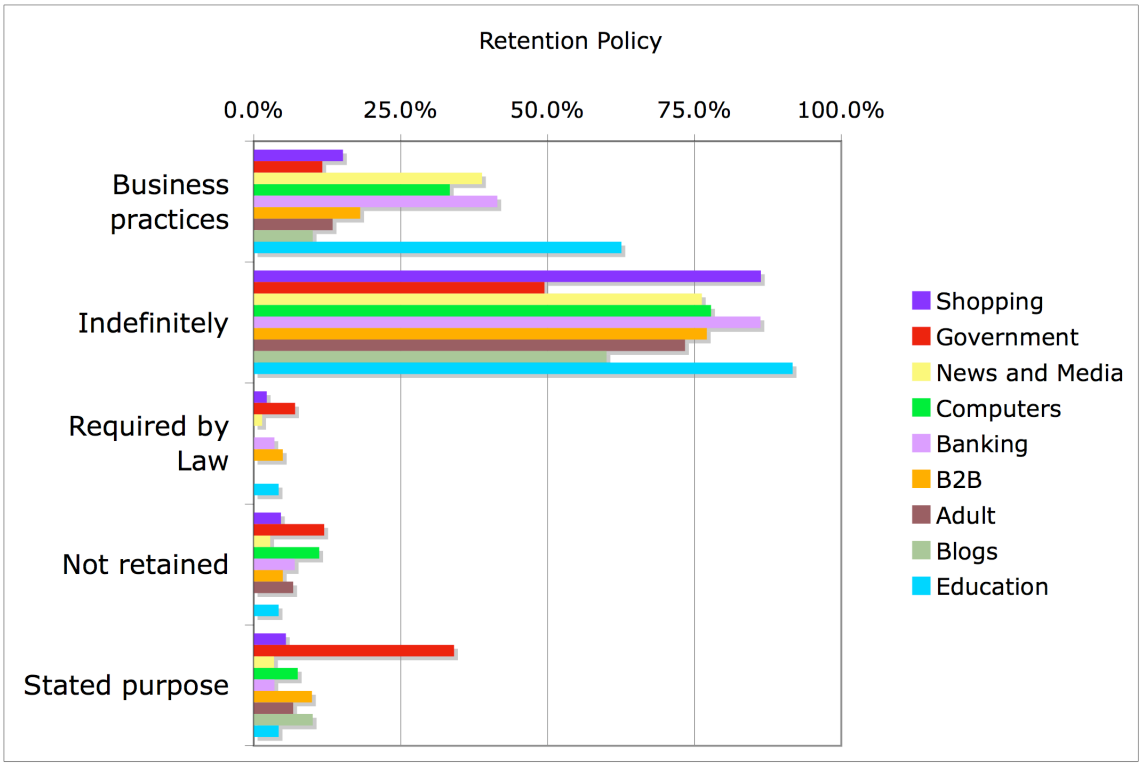

Figure 31: Data Retention Policy by Type of Website

5.4 Rate of Change of P3P Policies

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.

5.4.1 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 web sits. This set of 470 new policies includes 272 unique policies.

5.4.2 Policies Removed

During the study period 70 of the P3P policies that had been available at the beginning 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.

5.4.3 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 three 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.

5.5 P3P 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 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."[82] 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 website to FTC enforcement.

5.5.1 Syntactic Errors

When we attempted to validate the P3P policies we had collected, we found that the majority of these policies contained syntax errors. While a large number of policy errors were noted in the 2003 P3P study, our number is vastly greater.[83] In 2003, one third of the sites discovered contained errors as found by the W3C P3P Validator. However, when using the same validator with our study, we discovered that only 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.

Most of the errors in this study were considered "non-critical errors" in that they conflicted with the P3P specification, but at the same time the evaluator was still able to function correctly. These errors usually amounted to using an older version of the standard. This 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). The critical errors only accounted for about five percent of all of the URLs found; this number is similar to the 2003 statistic which found critical errors in six percent of the policies examined.

5.5.2 Types of Syntactic Errors

We used Perl code from the W3C's P3P validator as the basis for our own automated validator.[84] Using our validator, we were able to classify P3P syntax errors into the following fourteen categories:

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 Recommendation.

No policy name fragment — 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.

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.

Bad XML root — P3P policy or policy reference file has an invalid root node, which prevents the file from being parsed. This error prevents policy evaluation.

Policy expired — P3P policy or policy reference file has an explicit expiration date that is in the past.

Policy vocabulary error — P3P policy contains unrecognized elements or improperly references data elements. This error usually prevents policy evaluation.

No elements in policy — 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.

XML incorrectly formed — P3P policy or policy reference file are not valid XML documents.

Policy access error — P3P policy file cannot be accessed (HTTP 404, 403, etc.)

Namespace not specified — P3P policy or policy reference file does not include the P3P version number.

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.

No <META/> tag — Policy reference file does not begin with required <META/> tag.

No policy was found — A policy reference file exists but the P3P policy it references does not exist.

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.

Some of these errors are considered critical 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.

5.5.3 Errors in Popular Websites with P3P

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. Since becoming a W3C recommendation, 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, 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.

5.5.4 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 Popular sites.

Unfortunately, the sample size of P3P-enabled sites from the Popular List is too small to make any significant comparisons. However, it can be seen that the other P3P policies also suffer from the same common errors. Upon performing a z-test for proportions, we found that there was no significant difference with regard to the proportion of P3P policies containing old namespace versions. Nor was there a significant difference when comparing any of the other error categories, with two exceptions. 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 ($p < 0.0005$).

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

5.5.5 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.

5.5.6 Types of Semantic Errors

We used the sixty-seven APPEL files discussed in P3P Adoption Rates on page 34 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 12 shows these 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 12 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).

	<ACCESS> (1)	<CATEGORIES> (17)	<DISPUTES> (4)	<NON-IDENTIFIABLE> (1)	<PURPOSE> (12)	<RECIPIENT> (5)	<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	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 12: 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).

<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).

<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), nor do they collect any manner of contact information.

<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 web site's privacy practices.

<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.

<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 by 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.[85] 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.

<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.[86] Overall, 41 errors were made across the five different 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 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.

<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.

5.5.7 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.con 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 a certain 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.[87] The Financial Modernization Act of 1999, also known as the "Gramm-Leach-Bliley Act," requires institutions in the financial sector to publish privacy policies.[88] Wachovia, a bank, had some serious discrepancies between their P3P policy and their natural language policy. In Types of Syntactic Errors on page 63 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[89] 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. 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 than the one linked from the bottom of every page on the site.

5.5.8 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."

6 Discussion

One of the perennial arguments in online privacy is whether self-regulation is adequate. While we expect our readers have already formed considered opinions, we offer the following observations.

- The phrase "human readable privacy policy" may be a misnomer. Most people cannot use privacy policies because they cannot understand them. This problem has led to layered privacy policies (not studied here, but discussed by the FTC as one approach for financial policies[90]) and P3P. It remains an issue.
- The Gramm-Leach-Bliley Act contributed to policies that were more concrete, but still unreadable. Practices may have become slightly *less* privacy protective, in part as a result of mergers and acquisitions.
- Websites from nations within the European Union are more privacy protective than non-EU websites, perhaps as a result of EU legal requirements. This may be an interesting area for further research.

It appears legislation is neither a panacea nor innately evil. The best opportunity to affect change may be through contributing to better sample privacy policies.

P3P use continues to increase over time. P3P has reached a large number of countries and is associated with popular websites. However, people are not availing themselves of free tools to check the syntax of their P3P policies.

It appears that human readable policies simply omit mention of potentially objectionable information. For example, sites that keep information indefinitely or use it for telemarketing rarely tell customers of those practices. However, P3P policies offer greater specificity. As a result, P3P policies can lead to better informed consumers and greater consumer choice.

7 About Us

7.1 CyLab

Carnegie Mellon CyLab is a bold and visionary effort aimed at creating a public-private partnership to develop new technologies for measurable, available, secure, trustworthy, and sustainable computing and communications systems and to educate individuals at all levels.

Carnegie Mellon CyLab is a university-wide, multidisciplinary initiative involving more than 200 faculty, students, and staff at Carnegie Mellon that builds on more than two decades of Carnegie Mellon's leadership in Information Technology. CyLab works closely with the CERT® Coordination Center (CERT/CC), a leading, internationally recognized center of Internet security expertise. Through its connection to the CERT/CC, CyLab also works closely with US-CERT — a partnership between the Department of Homeland Security's National Cyber Security Division (NCSD) and the private sector — to protect our national information infrastructure. Please see http://www.cylab.cmu.edu/ for additional information.

7.2 CMU Usable Privacy and Security Laboratory

The CMU Usable Privacy and Security Laboratory (CUPS) was established in the Spring of 2004 to bring together Carnegie Mellon University researchers working on a diverse set of projects related to understanding and improving the usability of privacy and security software and systems. The privacy and security research community has become increasingly aware that usability problems severely impact the effectiveness of mechanisms designed to provide security and privacy in software systems. CUPS is affiliated with Carnegie Mellon CyLab. Please see http://cups.cs.cmu.edu/ for additional information.

7.3 CyLab Privacy Interest Group

The CyLab Privacy Interest Group (CPIG) is co-chaired by Jody Westby and Lorrie Cranor. Jody is a Distinguished Fellow with CyLab and President & CEO of Global Cyber Risk LLC. She also chairs the American Bar Association's Privacy & Computer Crime Committee. CPIG will serve as a forum for CyLab members to discuss technical, legal, and policy issues related to privacy, interact with the researchers involved in CyLab's privacy-related research initiatives, and provide input into future CyLab privacy activities.

7.4 Authors

Lorrie Faith Cranor is a CyLab faculty member and director of the CMU Usable Privacy and Security Laboratory (CUPS). She is an associate research professor in the School of Computer Science and in the Engineering & Public Policy department at Carnegie Mellon. She chaired the Platform for Privacy Preferences (P3P) Specification Working Group at the World Wide Web Consortium.

Serge Egelman is a doctoral student in the Computation, Organizations, and Society program within the School of Computer Science at Carnegie Mellon. His research centers on the usability of online security, privacy, and trust systems.

Aleecia M. McDonald is a doctoral student in Engineering & Public Policy at Carnegie Mellon. Her research interests include information technology policy and privacy issues.

Steve Sheng is a doctoral student in Engineering & Public Policy at Carnegie Mellon. His research interests include information technology policy and privacy, and online trust.

Appendix A: Top Websites

We used a list of 30,000 most frequently clicked on websites from AOL search data in October, 2005. In order to make our results comparable to other studies, we further refined the list by removing:

- all websites that had a top-level domain other than .com
- pornographic websites
- websites targeted to children

We also truncated to <hostname>.com and removed duplicates. For example, mail.yahoo.com and travel.yahoo.com became one instance of yahoo.com. All told, we needed 92 websites to generate our sample frame of 75.

The alphabetized list of websites we used follows. See "Comparison of Popular Websites to Random Websites" on page 7 for our analysis.

Underlined websites do not have human readable policies. Three of the top 75 sites have no privacy policies.

Highlighted websites have P3P policies. Twenty four of the top 75 sites have P3P policies. We compared the P3P encodings we created from human readable policies to the P3P policies the companies provided.

about.com	gamespot.com	rootsweb.com
aim.com	geocities.com	rottentomatoes.com
amazon.com	go.com	runescape.com
angelfire.com	google.com	rxlist.com
answers.com	hgtv.com	sing365.com
aol.com	homes.com	southwest.com
ask.com	hotmail.com	sportsline.com
askjeeves.com	imdb.com	superpages.com
azlyrics.com	infoplease.com	switchboard.com
bankofamerica.com	lyrics007.com	target.com
bizrate.com	lyricsdownload.com	ticketmaster.com
city-data.com	lyricsfreak.com	tripadvisor.com
cnn.com	lyricsondemand.com	tripod.com
cooks.com	mapquest.com	tv.com
dealtime.com	microsoft.com	ups.com
download.com	msn.com	usatoday.com
drugs.com	mtv.com	usps.com
ebay.com	myspace.com	wachovia.com
ehow.com	nextag.com	walmart.com
emedicine.com	nfl.com	washingtonpost.com
enchantedlearning.com	palottery.com	weather.com
epinions.com	partypop.com	wellsfargo.com
facebook.com	pogo.com	wunderground.com
findarticles.com	powerball.com	xanga.com
flalottery.com	qvc.com	yahoo.com

Appendix B: Random Websites

We selected 100 websites at random from the top 12,000 (out of 30,000) most frequently clicked on websites from AOL search data in October, 2005. In order to make our results comparable to other studies, we further refined the list by removing:

- all websites that had a top-level domain other than .com
- pornographic websites
- websites targeted to children

We also truncated to <hostname>.com and removed duplicates. For example, mail.yahoo.com and travel.yahoo.com became one instance of yahoo.com. All told, we needed 115 websites to generate our sample frame of 100.

The alphabetized list of websites we used follows. See "Comparison of Popular Websites to Random Websites" on page 7 for our analysis.

Underlined websites do not have human readable policies. Twenty two of the Random sites have no privacy policies.

Highlighted websites have P3P policies. Nine of the Random sites have P3P policies. We compared the P3P encodings we created from human readable policies to the P3P policies the companies provided.

007b.com	education-world.com	palmharbor.com
100bigcoupons.com	embassysuites.com	pantagraph.com
100megsfree3.com	feedsfarm.com	people.com
aa.com	fhm.com	poedecoder.com
aarphealthcare.com	funnyhub.com	priceviewer.com
abideinchrist.com	gamespot.com	protectmydna.com
about.com	hickerphoto.com	queenmary.com
adultchamber.com	hilton.com	rcuniverse.com
aldoshoes.com	hipusa.com	roxio.com
allrecipes.com	hmco.com	samsung.com
allsexycelebs.com	hurricanecity.com	shareup.com
americanhomeguides.com	ibc.com	siouxcityjournal.com
ascap.com	intermatic.com	skateboarding.com
assist2sell.com	itsmarta.com	socialservice.com
atptennis.com	jackdaniels.com	spirithome.com
audreysmotherofthebride.com	kindredtrails.com	stltoday.com
baby-parenting.com	kirotv.com	suzuki.com
baby-place.com	kitchengifts.com	systransoft.com
bobevans.com	lifewaystores.com	thefunplace.com
candylandcrafts.com	lycos.com	trains.com
celebsmovies-online.com	majorityreportradio.com	untied.com
citysearch.com	mazdausa.com	usedwreckingyards.com
cjonline.com	molottery.com	vanityfair.com
coolquiz.com	movieeye.com	vibe.com
costplus.com	mrfreefree.com	wcbs880.com
creatingkeepsakes.com	mtve.com	webleaguemanager.com
cubcadet.com	myschoolonline.com	websterbank.com
daysofourlives.com	mysteries-megasite.com	weddinggazette.com
democratandchronicle.com	ncchildsupport.com	whitneybank.com
denverbroncos.com	nestofdeath.com	wolfram.com
drahoshcreations.com	neteller.com	worldofwatches.com
dresses.com	nhl.com	yahoo.com
drivewire.com	nyse.com	
e-cards.com	pacode.com	

Appendix C: P3P Elements

The table below provides the recommended human-readable translation of P3P elements, as provided by the P3P specification.[91]

Access

nonident	We do not keep any information identified with you
all	We give you access to all of our information identified with you
contact-and-other	We give you access to your contact information and some of our other information identified with you
ident-contact	We give you access to only your contact information in our records
other-ident	We allow you to access some of our information identified with you, but not your contact information
none	We do not give you access to our information about you

Disputes

service	[display long description and short description, if provided, with hyperlink to service URI, otherwise display "customer service" with hyperlink to service URI]
independent	[display long description and short description, if provided, with hyperlink to service URI, otherwise display "independent organization" with hyperlink to service URI]
court	We believe that the following authority offers recourse for disputes: [display long description and short description, if provided, with hyperlink to service URI, otherwise display "possible legal complaint" with hyperlink to service URI]
law	We believe that the following laws or regulations provide recourse: [display long description and short description, if provided, with hyperlink to service URI, otherwise display "law" with hyperlink to service URI]

Purpose

current	To provide the service you requested
admin	To perform website and system administration
develop	For research and development, but without connecting any information to you
tailoring	To customize the site for your current visit only
pseudo-analysis	To do research and analysis in which your information may be linked to an ID code but not to your personal identity
pseudo-decision	To make decisions that directly affect you without identifying you, for example to display content or ads based on links you clicked on previously
individual-analysis	To do research and analysis that uses information about you
individual-decision	To make decisions that directly affect you using information about you, for example to recommend products or services based on your previous purchases
contact	To contact you through means other than telephone (for example, email or postal mail) to market services or products
historical	To aid in historical preservation as governed by a law or policy described in this privacy policy
telemarketing	To contact you by telephone to market services or products
other-purpose	For other uses: [include site's human, readable explanation; if site omits human-readable explanation say "not described here"]
required (attribute of purpose and recipients elements)	(attribute, see below)
required always	(no remark)
required opt-in	[append to purpose/recipient] -- only if you request this
required opt-out	[append to purpose/recipient] -- unless you opt-out

Recipient

ours	Companies that help us fulfill your requests (for example, shipping a product to you), but these companies must not use your information for any other purpose
delivery	Delivery companies that help us fulfill your requests and who may also use your information in other ways
same	Companies that have privacy policies similar to ours
other-recipient	Companies that are accountable to us, though their privacy policies may be different from ours
unrelated	Other companies whose privacy policies are unknown to us
public	People who may access your information from a public area, such as a bulletin board, chat room, or directory
required	(attribute of purpose and recipients elements) (attribute, see below)
required always	(no remark)
required opt-in	[append to purpose/recipient] -- only if you request this
required opt-out	[append to purpose/recipient] -- unless you opt-out

Retention

no-retention	We do not keep your information beyond your current online session
stated-purpose	We keep your information only long enough to perform the activity for which we collected it
legal-requirement	We keep your information only as long as we need to for legal purposes
business-practices	Our full privacy policy explains how long we keep your information
indefinitely	We may keep your information indefinitely

Categories

physical	Name, address, phone number, or other physical contact information
online	Email address or other online contact information
uniqueid	Website login IDs and other identifiers (excluding government IDs and financial account numbers)
purchase	Information about your purchases, including payment methods
financial	Financial information such as accounts, balances, and transaction history
computer	Information about the computer you are using, such as its hardware, software, or Internet address
navigation	Which pages you visited on this website and how long you stayed at each page
interactive	Activities you engaged in at this website, such as your searches and transactions
demographic	Information about social and economic categories that might apply to you, such as your gender, age, income, or where you are from
content	Messages you send to us or post on this site, such as email, bulletin board postings, or chat room conversations
state	Cookies and mechanisms that perform similar functions
political	Which groups you might be a member of such as religious organizations, trade unions, and political parties
health	Health information such as information about your medical condition or your interest in health-related topics, services, or products
preference	Information about your tastes or interests
location	Information about an exact geographic location, such as data transmitted by your GPS-enabled device
government	Government-issued identifiers such as social security numbers
other-category	Other types of data: [include site's human, readable explanation; if site omits human-readable explanation say "not described here"]

End Notes

[1] Lorrie Faith Cranor, Joseph Reagle and Mark S. Ackerman, "Beyond Concern: Understanding Net Users' Attitudes About Online Privacy" (AT&T Labs-Research, 1999); Privacy Leadership Initiative, "Consumer Privacy Attitudes and Behaviors Survey" (New York City: 2001); "Privacy Leadership Initiative, A Survey of Consumer Privacy Attitudes and Behavior" Conducted by Harris Interactive (2000); Louis Harris and Associates and Alan F. Westin, "E-Commerce Privacy Survey" (Privacy & American Business and Price Waterhouse, Inc., 1998); Joseph Turow, "Americans and Online Privacy: The System Is Broken" (Philadelphia: Annenberg Public Policy Center, University of Pennsylvania, 2003).
[2] Robert Lemos, "Data security moves front and center in 2005," *SecurityFocus,* December 29, 2005, http://www.securityfocus.com/news/11366, accessed 20 October, 2006.
[3] IBID.
[4] Robert Lemos, "Veterans Affairs Warns of Massive Privacy Breach," *SecurityFocus*, May 22, 2006. http://www.securityfocus.com/news/11393, accessed 21 October, 2006.
[5] "A Chronology of Breaches Since the ChoicePoint Incident," http://www.privacyrights.org/ar/ChronDataBreaches.htm, accessed 19 October 2006.
[6] "Spying Scandal at Hewlett-Packard," *CSOonline*, http://www2.csoonline.com/exclusives/column.html?CID=25047 Accessed 19 October, 2006.
[7] IBID.
[8] Elise Ackerman, "After privacy breach, AOL executive resigns," August 21, 2006, *San Jose Mercury News.* http://www.mercurynews.com/mld/mercurynews/business/15328282.htm Accessed 19 October 2006.
[9] "Ready for Some Spear Phishing?" *IT Business Edge*, September 22, 2006, http://www.itbusinessedge.com/item/?ci=20508 Accessed 20 October 2006.
[10] Bill Brenner, "Data Storage Bills Go to Extremes," May 12, 2006, *SearchSecurity*, http://searchsecurity.techtarget.com/columnItem/0,294698,sid14_gci1188020,00.html, accessed 21 October 2006.
[11] Microsoft press release, "Microsoft Advocates Comprehensive Federal Privacy Legislation," November 3, 2005. http://www.microsoft.com/presspass/press/2005/nov05/11-03DataPrivacyPR.mspx Accessed 26 January 2007.
[12] Brian Krebs, "Microsoft Calls for National Privacy Law," *Washington Post*, November 3, 2005

http://blog.washingtonpost.com/securityfix/2005/11/microsoft_calls_for_national_p_1.html Accessed 26 January 2007.
[13] Available for download from http://www.microsoft.com/downloads/details.aspx?FamilyId=C48CF80F-6E87-48F5-83EC-A18D1AD2FC1F&displaylang=en as of 21 October 2006.
[14] Microsoft Corporation, *Privacy Guidelines for Developing Software Products and Services*, Version 2.1, October 10, 2006.
[15] Paul Festa, "Judge Tosses Online Privacy Case," *CNET News.com*, June 16, 2004, http://news.com.com/Privacy+advocates+protest+Northwest+dismissal/2100-1023_3-5234971.html Accessed 21 October 2006.
[16] IBID.
[17] David A. Stampley, "Managing information technology security and privacy compliance," *Neohapsis* May 29 2005, http://www.neohapsis.com/utility/NeoPrivacyWhitepaper.pdf (linked to as "Privacy Compliance") Accessed 31 October, 2006.
[18] IBID.
[19] IBID.
[20] IBID.
[21] KPMG provided the list of top 500 institutions.
[22] The Internet Archive is a service that allows people to visit archived versions of Websites from 1996 to the present time. See http://www.archive.org/ , last accessed March 12, 2006.
[23] http://privacyfinder.org/
[24] Google, Inc. Froogle, 2005. http://froogle.google.com/.
[25] Simon Byers, Lorrie Faith Cranor, and David Kormann. Automated Analysis of P3P-Enabled Websites. In *Proceedings of the Fifth International Conference on Electronic Commerce* (ICEC2003). Pittsburgh, PA, October 1-3, 2003.
[26] Privacy Bird http://www.privacybird.com/ Accessed 4 November 2006.
[27] Simon Byers, Lorrie Faith Cranor, and David Kormann. Automated Analysis of P3P-Enabled Websites. In *Proceedings of the Fifth International Conference on Electronic Commerce* (ICEC2003). Pittsburgh, PA, October 1-3, 2003.
[28] Serge Egelman, Lorrie Faith Cranor, and Abdur Chowdhury. An Analysis of P3P-Enabled Websites among Top-20 Search Results. *Proceedings of the Eighth International Conference on Electronic Commerce* August 14-16, 2006, Fredericton, New Brunswick, Canada.
[29] Federal Trade Commission, *Privacy Online: A Report to Congress* (Federal Trade Commission, 1998); Federal Trade Commission, *Privacy Online: Fair Information Practices in the Marketplace: A Report to Congress* (Washington DC: Federal Trade Commission, 2000).

[30] Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, Office of the Comptroller of the Currency and Office of Thrift Supervision, *Interagency Financial Institution Website Privacy Survey Report* (Washington DC: 1999).
[31] George R. Milne and Mary Culnan, "Using the Content of Online Privacy Notices to Inform Public Policy: A Longitudinal Analysis of the 1998-2001 U.S. Web Surveys," *The Information Society* 18(5), 2002.
[32] The Platform for Privacy Preferences 1.1 (P3P1.1) Specification, http://www.w3.org/TR/P3P11/ Accessed 6 November 2006.
[33] IBID.
[34] The P3P specification also includes the Historical purpose. We have not indicated it here because it applies mostly to government websites and we discovered late in the editing process that our students had not properly coded it.
[35] Federal Trade Commission, "Fair Information Practice Principles," http://www.ftc.gov/reports/privacy3/fairinfo.htm Accessed 29 November 2006.
[36] Federal Trade Commission, "Final Report of the FTC Advisory Committee on Online Access and Security," May 15, 2000.
http://www.ftc.gov/acoas/papers/finalreport.htm Accessed 30 January 2007.
[37] www.truste.org/
[38] www.bbbonline.org/
[39] Readability Info, "Readability Grades," http://readability.info/info.shtml, Accessed 6 November 2006.
[40] Readability Info, "Common Readability Scores," http://readability.info/commonscores.shtml, Accessed 6 November 2006.
[41] Xinguang Sheng and Lorrie Faith Cranor, "An Evaluation of the Effect of US Financial Privacy Legislation Through the Analysis of Privacy Policies," *I/S: A Journal of Law and Policy for the Information Society* 2 (3), Fall 2006, p. 943-979.
[42] Turkington and Allen, *Privacy Law: Cases and Materials.*
[43] The FTC and eight other federal agencies, charged to implement the GLB privacy rule, came up with a set of guidelines to implement the GLB (16 CFR Part 313).
[44] Other exceptions that allow for data disclosure without notice and choice include: (1) disclosures in response to consumer requests for specified financial services, (2) when necessary to protect the financial institution's records and to prevent actual or potential fraud, (3) disclosures to insurance rate advisory organizations, (4) disclosures to law enforcement agencies, (5) disclosures in connection with proposed sales or mergers, (6) disclosures to the financial institution's lawyers and accountants and (7) in compliance with Federal, State or local legal requirements. These exceptions are arguably in alignment with customer expectations of privacy. Thus, in this paper, we will not address these exceptions.

[45] "Affiliate is defined as any company that controls, is controlled by or under the common control with another company." (16 CFR 313.3 a) As an illustration, Citigroup owns Citibank, Dinner Club Card, and Smith Barney Investment. These companies can be considered affiliates. (Source: http://www.citigroup.com/citigroup/business/brands.htm)
[46] Virginia Boyd, "Financial Privacy in the United States and the European Union: A Path to Trans-Atlantic Regulatory Harmonization," (Harvard Law School, 2005).
[47] R.C. Turkington and A. L. Allen, *Privacy Law: Cases and Materials*. Second edition American Casebook Series. 2002.
[48] Neal R. Pandozzi, "Be Ware of Banks Bearing Gifts: Gramm-Leach-Bliley and the Constitutionality of Federal Financial Privacy Legislation," *Miami Law Review* 163 (2001).
[49] Paul M. Schwartz, "Property, Privacy, and Personal Data," *Harvard Law Review* 117 Harv. L. Rev. 2055 (2004).
[50] Peter P. Swire, "The Surprising Virtues of the New Financial Privacy Law," *Minnesota Law Review* 86. Minnesota Law Review.
[51] The Internet Archive is a service that allows people to visit archived versions of Websites from 1996 to the present time. See http://www.archive.org/ , last accessed March 12, 2006.
[52] As part of the coding process, the coders read each privacy policy carefully to determine if the policy mentioned anything about whether the company shares information with affiliates or third parties, and if so what information is shared, and what choice is given to consumers. If after reading the privacy policy, the coders could not determine the answers to these questions, they coded it as "unknown."
[53] Data for the top 10 group: affiliate sharing unknown: pre-GLB, 20%, post-GLB 0%; affiliate disclosure: pre-GLB, 30%, post-GLB 0%; affiliate choice: pre-GLB, 30%, post-GLB 0%; Third party sharing unknown: pre-GLB, 10%, post-GLB 0%; third party choice unknown: pre-GLB, 40%, post GLB 0%.
[54] We have calculated various other readability measures such as ARI, and Gunning Fog index, and found similar trends.
[55] U.S Census Bureau, Educational Attainment in the United States: 2004 , Table 1. available at http://www.census.gov/population/www/socdemo/education/cps2004.html, accessed March 26, 2006
[56] Available at http://www.aba.com/About+ABA/ABA_privprinpublic.htm.
[57] Boyd, "Financial Privacy in the United States and the European Union: A Path to Trans-Atlantic Regulatory Harmonization."
[58] W3C Working Group, *The Platform for Privacy Preferences 1.1 (P3P1.1) Specification*, updated November 13, 2006, http://www.w3.org/TR/P3P11/ Accessed 30 January, 2007.

[59] "Platform for Internet Content Selection (PICS,)" *W3C*, updated July 8, 2005, http://www.w3.org/PICS/ Accessed 23 October, 2006.
[60] Hal Abelson, Mike Fischer, and Joanne Costello, "The Internet Censorship Saga: 1994-1997," modified October 6, 1999, http://www.swiss.ai.mit.edu/6095/articles/cda/saga.html Accessed 23 October, 2006.
[61] Lorrie Faith Cranor, *Web Privacy with P3P* (Sebastopol: O'Reilly and Associates, 2002) page 44.
[62] IBID. See chapter 4 for a full discussion of P3P's development.
[63] Serge Egelman, Lorrie Faith Cranor, and Abdur Chowdhury. An Analysis of P3P-Enabled Websites among Top-20 Search Results. *Proceedings of the Eighth International Conference on Electronic Commerce* August 14-16, 2006, Fredericton, New Brunswick, Canada.
[64] E. Hargittai. The Changing Online Landscape: From Free-for-All To Commercial Gatekeeping. *Community Practice in the Network Society: Local Actions/Global Interaction*, pages 66–76, 2004. http://www.eszter.com/research/c03-onlinelandscape.html.
[65] D. Fellows. Search Engine Users. January 23, 2005. http://www.pewinternet.org/pdfs/PIPSearchengineusers.pdf.
E. Burns. Search increased in august, October 7, 2005. http://www.clickz.com/stats/sectors/search tools/article.php/3554731.
[66] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th World Wide Web Conference,* 1998. http://www-db.stanford.edu/pub/papers/google.pdf.
[67] 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.
[68] S. Byers, L. F. Cranor, and D. Kormann. Automated Analysis of P3P-Enabled Websites. In *Proceedings of the Fifth International Conference on Electronic Commerce* (ICEC2003), October 1-3, 2003. http://lorrie.cranor.org/pubs/icec03.html.
[69] W. Adkinson, J. Eisenbach, and T. Lenard. Privacy online: A report on the information practices and policies of commercial web sites. Technical report, Progress & Freedom Foundation, 2002. http://www.pff.org/publications/privacyonlinefinalael.pdf.
[70] Ernst & Young. P3P Dashboard Report. August 2002. http://www.ey.com/global/download.nsf/US/P3P Dashboard - August 2002/$file/P3PDashboardAugust2002.pdf.
Ernst & Young. P3P Dashboard Report. January 2003. http://www.ey.com/global/download.nsf/US/P3P Dashboard - January 2003/$file/E&YP3PDashboardJan2003.pdf.

[71] Google, Inc. Froogle, 2005. http://froogle.google.com/.
[72] Office of Management and Budget. About E-GOV, 2005. http://www.whitehouse.gov/omb/egov/g-4-act.html.
[73] Mark Ward, "Net gains for Tuvalu," *BBC News* December 12, 2000, http://news.bbc.co.uk/1/hi/sci/tech/1067065.stm Accessed 5 November 2006. [NB: The article incorrectly claims Idealab is a Canadian company; it is a United States corporation based in Pasadena, California.]
[74] CIA, *The World Factbook – Cocos (Keeling) Islands*. https://www.cia.gov/cia/publications/factbook/geos/ck.html Accessed 3 November, 2006.
[75] VeriSign, "eNIC and VeriSign Partner to Expand Reach of .cc Domain," http://www.verisign.com/verisign-inc/news-and-events/news-archive/us-news-2001/page_200312171524283.html Accessed 5 November 2006.
[76] Europa, October 2005, "Directive 95/46/EC of the European Parliament," http://europa.eu.int/smartapi/cgi/sga_doc?smartapi!celexapi!prod!CELEXnumdoc&lg=EN&numdoc=31995L0046&model=guichett Accessed 5 November 2006.
[77] For more information, please see http://www.truste.org/
[78] For more information, please see http://www.bbbonline.org/privacy/
[79] dir.yahoo.com
[80] For more information, please see http://www.truste.org/
[81] For more information, please see http://www.bbbonline.org/privacy/
[82] 15 U.S.C. §45(a)
[83] S. Byers, L. F. Cranor, and D. Kormann. Automated Analysis of P3P-Enabled Websites. In *Proceedings of the Fifth International* Conference on Electronic Commerce (ICEC2003), October 1-3, 2003. http://lorrie.cranor.org/pubs/icec03.html.
[84] Y. Koike and S. Taiki. P3P Validator, January 29, 2002. http://www.w3.org/P3P/validator.html.
[85] Cranor provides advice on distinguishing these purposes on p. 94-95 of *Web Privacy with P3P*.
[86] L. F. Cranor, P. Guduru, and M. Arjula. User interface for privacy agents. *ACM Transactions on Computer-Human Interaction*, 13(2), June, 2006. http://portal.acm.org/citation.cfm?doid=1165734.1165735.
[87] 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 web sites generally post. Whether a policy is in a machine-readable code that is translated by a user agent, or simply in HTML on a web site, the policy constitutes a representation to consumers on which they can be expected to rely"

Source: L. Cranor and D. Weitzner. Summary report - w3c workshop on the future of p3p. Technical report, World Wide Web Consortium, November 2002. http://www.w3.org/2002/12/18-p3p-workshop-report.html.

[88] 15 U.S.C. §6801 et seq.

[89] http://www.wachovia.com/inside/legal footer/0,,2157 2158,00.html

[90] Letter from Hunton & Williams to the Federal Trade Commission, October 28, 2004, www.ftc.gov/os/comments/prescreenedoptout/OL-100022.pdf

[91] The Platform for Privacy Preferences 1.1 (P3P1.1) Specification, http://www.w3.org/TR/P3P11/ Accessed 6 November 2006.