Tracking Website Data-Collection and Privacy Practices with the iWatch Web Crawler

Carlos Jensen
School of EECS
Oregon State University
Corvallis, OR 97331, USA
+1-541-737-2555
cjensen@eecs.osrt.edu

Chandan Sarkar
School of EECS
Oregon State University
Corvallis, OR 97331, USA
sarkar@eecs.orst.edu

Christian Jensen
Department of Economics
Southern Methodist Uni.
Dallas, TX 75275, USA
christia@mail.smu.edu

Colin Potts
College of Computing
Georgia Institute of Tech.
Atlanta, GA 30332, USA
potts@cc.gatech.edu

ABSTRACT

In this paper we introduce the iWatch web crawler, a tool designed to catalogue and analyze online data practices and the use of privacy related indicators and technologies. Our goal in developing iWatch was to make possible a new type of analysis of trends, the impact of legislation on practices, and geographic and social differences online. In this paper we present preliminary findings from two sets of data collected 15 months apart and analyzed with this tool. Our combined samples included more than 240,000 pages from over 24,000 domains and 47 different countries. In addition to providing useful and needed data on the state of online data practices, we show that iWatch is a promising approach to the study of the web ecosystem.

Categories and Subject Descriptors

H.5.4 **[Hypertext/Hypermedia]**, Architecture, User issues
K.4.1 **[Public Policy issues]**. Privacy, Transborder data flow.

General Terms

Management, Measurement, Documentation, Human Factors, Standardization, Legal Aspects, Verification.

Keywords

Privacy, Demographics, Data-collection practices, Web-crawling, Cookies, Webbugs, P3P, Legislative impact.

1. INTRODUCTION

The Web is a complex place in terms of technologies and practices, especially when considering how these affect privacy and security. These are important concerns for consumers who have to decide who to trust with their data, for legislators who have to develop meaningful and effective regulation, as well as for system administrators and developers, who stand to loose significant time and money on flawed models and designs, or potentially face a user backlash and/or fines.

Part of what makes this such a challenging problem is that technology and business practices are constantly evolving. Keeping up with changes and trends can sometime seem like a full-time job. Another challenge is that the web is a global system, crossing and blurring many of the traditional lines of jurisdictions. A company can be registered in one country, be hosted in a number of other countries, and do business with consumers from anywhere in the world. This picture can get even more complicated when we start talking about multi-national companies, and potential business-to-business (b2b) partners. This issue of jurisdiction has been, and will continue to be for the foreseeable future, a serious challenge to e-commerce and e-business. Determining compliance should therefore be a major concern for designers, developers, and administrators of such systems.

For legislators and policy makers it is therefore important to understand the impact of policy decisions in order to craft rules and legislation which will be effective and meaningful, and enforce such rules once adopted. Given that legislation often lags behind technological adoption and development, it is important to monitor when safeguards are needed, and when they are no longer meaningful or necessary. It is equally important to monitor developments following the introduction of new legislation as well, to ensure that these are having the intended and desired effects, something which is not always the case [22].

For consumers it is important to understand the risks out there - including the prevalence of undesirable or dubious security and privacy practices - in order to make better decisions about whom to trust. This is especially important as a mechanism for ensuring market forces take effect. If consumers are unaware of companies using undesirable practices, they cannot express their preferences by taking their business elsewhere. Such knowledge can help spur the adoption of effective and necessary safeguards and detection mechanisms, and can help end-users press legislators for regulation of practices.

For researchers, it is important to know what problems, technologies and practices are worth addressing, or which remedies are having effect. When designing monitoring, notification, blocking, or any other type of technologies, it is important to know where best to invest time and effort, especially given the limited resources in many academic settings. Such an overview could help researchers make the necessary justifications for their decisions.

In order to meet the information needs of such diverse stakeholders we need access to a reliable set of data about current data practices and technology use. Because this data may influence public policy, consumer perception, as well as business practices, it is essential that the data be publicly available, and collected in a transparent and unbiased fashion. A technique for doing this is to instrument a web crawler, specifically designed to

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

go out and index web-pages based on publicly visible and machine identifiable data-collection practices and policies. This data could then be made available to the public, and/or scrutinized, and used as a common benchmark or reference set. This basic approach has been used in the past [10], though not on the scale of what we demonstrate in this paper.

Our proposal for filling this function is a web crawler named iWatch. The name is derived from the famous question “Q*uis custodiet ipsos custodes?*” or "*Who watches the watchers/guards?*" originally posed by Plato in *The Republic* [31] and popularized in Latin in Juvenal’s *Satires* [24]. In this case, iWatch monitors those who normally monitor us; websites.

iWatch is meant to serve as a source of basic statistics on the state of privacy, security, and data-collection practices on the web. Because we have no access to information on what websites are doing behind the scenes we have to limit our analysis to the information and technologies which are publicly visible, and what we can automatically detect and analyze. Though this naturally limits the accuracy and scope of our analysis, it still allows us to examine and detect some fairly interesting practices and situations.

In this paper we set out to demonstrate the feasibility and value of this approach to analyzing real-world data-practices from the perspective of the outside observer (no knowledge of internal website workings). We will look at several interesting practices, and ways of examining the data. This paper is also meant to serve as a point for reflection and discussion about which practices to observe, and how the raw data from a system such as iWatch, which is still a work in progress, can and should be evolved and made available to a wider audience.

The structure of the rest of this paper will be as follows: We will first discuss a selection of related work, followed by a description of the terminology, conventions and definitions used in this paper. We then discuss the workings and implementation decisions made in our web-crawler, and present two sets of data, from 2005, and 2006, and explore the changes which have taken place in this period, as well as the impact of geography and regulation. We wrap up with a discussion of these results and future plans.

2. RELATED WORK

Privacy and security have long been recognized as important areas of concern, both offline and online. As such, this is one of the areas where online activity already has a long history of legislation. These laws have taken different forms across the globe. In Europe, comprehensive or omnibus laws for data protection have been enacted, while the US has largely implemented sector specific laws. These two approaches are fundamentally different, both approaches having advantages and disadvantages, which are often hotly debated [26, 33].

Regardless of approach, the goal of these privacy laws is to protect the Personally Identifiable Information (PII) of the individual, as well as regulate how information may be collected, for what purpose, and how it must be protected. Examples of such laws include the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) [36], the US Children’s Online Privacy Protection Act of 1998 (COPPA) [34], the US Gramm-Leach-Bliley Financial Services Modernization Act of 1999 (GLBA) [35], and the European Union Directive on the protection of personal data (95/46/EC) [17].

Given that studies have shown that users fail to read sites’ privacy policies [22, 27], the kinds of minimum protections these laws put in place are particularly important. Previous research has shown that legislation can have mixed effects on policies, especially their readability and usability [3, 22].

Despite legislative efforts, privacy concerns have been shown to be major obstacles to the adoption and success of e-commerce [1, 23]. Numerous surveys indicate that people consider privacy to be important [6, 7, 11, 15]. Privacy concerns are the most cited reasons for avoiding the use of e-commerce systems, an aversion that industry groups estimate costs e-commerce companies USD 25 billion per year in lost revenue opportunities [23]. Surveys have also found that people are more concerned about their privacy online than offline [21], even though most cases of identity theft occur offline [20]. It is not surprising that industry groups invest significant resources to build consumer confidence and engage in voluntary efforts such as publishing privacy policies and seeking different forms of certification.

Such self-regulation attempts through programs such as seal programs such as TRUSTe (http://www.truste.org), BBBOnLine (Better Business Bureaus Online Seal, http://www.bbbonline.org), MultiCheck and WebTrust (offered by American Institute of CPAs http://www.cpawebtrust.org) allow licensees who abide by posted privacy policies and/or allow compliance monitoring to display the granting organization’s seal of approval on their web site. Such programs have been found to significantly increase consumer trust [21, 28, 29], though some questions remain over whether what they imply matches user expectations, and questions remain about the ease with which sites may misrepresent their certification status [29]. In other words, there is some indication that users are being misled, intentionally or unintentionally, by some of these efforts [27].

We also know from surveys that though users think it is important for sites to present privacy policies, they are less than impressed with their quality and accuracy [12]. Surveys show that users find privacy policies to be boring, hard to read and understand, hard to find, and that they don’t answer the kinds of questions they are interested in. The same survey also found that most people do not believe the claims and guarantees made in privacy policies [12, 20]. While most surveys report that a sizable portion of users claim to read such policies or notices regularly [12], there is evidence to suggest these reports are greatly exaggerated [21].

To overcome some of the problems associated with privacy policies and reduce the burden on users, machine-readable policy specification languages, such as P3P [8, 9] and EPAL [5], have been proposed. These policies can be read by automated agents (such as Privacybird [9], Privacy Fox [4], or the Microsoft IE 6 and 7, or Netscape 7 browsers themselves), only alerting users if the policy is likely to cause concern. The theory is that by filtering out the noise and drawing users’ attention to only those policy elements which require attention, users are more likely to be engaged.

The most popular and widely used of these technologies without question is P3P. The Platform for Privacy Preferences (P3P) was created by the W3C to make it easier for web site visitors to obtain information about sites’ privacy policies [8, 9]. P3P specifies a standard XML format for machine-readable privacy policies that can be parsed by a user-agent program. These tools have shown some indications of success [16], though there is little data on their effects during long-term or large-scale use. P3P policies have also been used as data to direct users’ web-searches

[10] in a system sharing many methodological similarities to our iWatch.

A number of other tools independent of P3P have also been developed over the years, including filtering and privacy protecting proxy servers, popup-blockers, cookie blockers and analysis tools, anti-phishing tools, etc. Given that many of these functions have subsequently been absorbed by the latest generation of web-browsers, their numbers and user base is unknown today.

Regardless of the underlying technology, HCI researchers have been examining the issue of how to improve the usability and usefulness of such systems, an early shortcoming of many. Classic papers and studies include [37, 38]. This research showed that a secure system would fail unless these security measures were made usable. In recent years we have seen excellent papers on why phishing attacks work [14, 13], and how our tools and warning tend to go unheeded, regardless of the information presented [39]. While excellent results, it is obvious more work still needs to be done in this area as there are far more studies of why things fail than how to succeed.

Our approach of harvesting and examining large amounts of data via the use of a web-crawler has been employed by other security and privacy researchers. Recently, this approach has produced interesting results in the identification of malware and spyware disseminating websites [30, 32]. In these studies, researchers were able to scan and classify a large enough sample to convincingly argue about the state of the Internet as a whole.

3. Definitions

Before diving into the meat of our study, it is important to define certain terms in order to avoid misunderstandings or ambiguity. Our definitions should most often match generally accepted definitions, but may in some cases have a rather more narrow definition, chosen for practical considerations.

In this paper, domain, web server, and website are terms which are used interchangeably. While in the real-world, a given domain can host many distinct sites, we differentiate between sites based solely on domain-names. A distinct domain-name in our study identifies a distinct domain. Our classification of domains was very simplistic. We did not attempt to identify synonymous domain names (www.theregister.co.uk is not recognized as a synonym for www.theregister.com), or sub-domains (news.bbc.co.uk is not identified as a sub-domain of www.bbc.co.uk). The first is a hard problem and requires either a set of records from domain registrars, or a lot of hand-tuning. The second, though technically simple to implement, would cause problems with hosting services and smaller or related web-sites, which may lack unique second-level domain names.

We will also use the terms 1st party and 3rd party frequently. In this context a 1st party typically refers to the domain or website which served the page, and a 3rd party is any other domain/website which either receives information about the transaction, or supplies information or resources used by the requested page. Examples are 3rd party cookies, webbugs, and banner ads.

In this paper we will talk about technologies such as P3P policies, webbugs, cookies, popups, and banners. P3P stands for the Platform for Privacy Preferences, and is a standard for specifying privacy policies in a machine-readable XML format [8]. There are two types of P3P policies, the compact policy (CP) and the full policy. The P3P compact policy is a keyword abbreviated P3P policy, offering less detail and nuance, but often used by browsers to filter cookies. P3P and P3P policy will be terms that are used interchangeably in this paper.

The P3P protocol specifies 3 ways of publishing a P3P policy; in the HTTP header (can either be a compact policy, or a link to a full policy), in the HTML document as a link tag, or in a well-known location on the server. Because of some quirks of the way web servers implement the serving of P3P policies (see discussion in methodology), our current version of iWatch only finds policies posted in the HTTP header or the body of the document, it does not search the known locations. In order to fetch these remaining policies without bringing the crawler to a halt we delegate this task to a standalone program.

Privacy Seals are, in this paper, a combination of different certificates or trustmarks issued by TRUSTe and BBBOnline (BBBPrivacy and BBBReliability seals). These seals certify that the site discloses or follows a minimum set of privacy protection and security practices. While different seals or certificates are enforced by different agencies, have different meanings, and offer different enforcement mechanisms and guarantees, they are all meant to calm potential users concerns. Given the relatively low usage numbers, the different seal programs are grouped together for most of our analysis.

Webbugs, also known as web-beacons or pixel tag, are a collection of techniques aimed to tag and collect information from web and email users without their knowledge. In a web page, webbugs are typically used to track users navigating a given site, and have become quite ubiquitous. Webbugs technically can be implemented through a number of different techniques, but are most commonly associated with a 1x1 pixel transparent gif, invisible to the user. Webbugs are often used to augment the tracking available with cookies, and are most troubling when set by third parties, usually without user knowledge or consent. In iWatch we group a number of tracking techniques under the label of webbugs, but only when these are set and used by 3rd parties. We do not classify banner ads or 3rd party cookies as webbugs, but rather track these separately.

Much has been written about cookies, and so a discussion of how they work and their potential threats to user privacy is omitted here. We will just mention that in this work we do track the three main categories of cookies separately, session cookies, defined as cookies set by the first party and expiring with the browsing session, 1st party cookies, set by the 1st party and set to persist, and 3rd party cookies, which are set for any domain other than the 1st party.

Unsolicited popups, or just popups for short, refers to the much hated technique of opening new browser windows, typically for the purpose of advertising. Affiliated techniques include the pop-under (popups which try to hide themselves). They present a potential danger to end-users as they often serve up content for third parties, enabling these to track users much like webbugs. Popups have stopped being as big a focus in recent years as blocking tools and techniques have become ubiquitous and effective.

Web banners, or banners for short, do not present a privacy risk in and of themselves, unless served by a third party. In this case, they serve much the same function as a webbug, though at least remaining visible to the user. Banners in our study are identified by their size (these are the standardized sizes set by the Internet

Advertising Bureau (http://www.iab.net/standards/adunits.asp)), and the fact they are served by a 3rd party.

Some practices and technologies are ambiguous or difficult to detect reliably. This is especially true for automatic pop-ups, which at times are difficult to disambiguate from user-activated pop-ups, or webbugs from images or tricks used to layout web-pages. While we have done our best to unambiguously define and detect interesting practices, there is still room for improvement. Webbugs and unsolicited popups are still difficult to detect unambiguously, and some amounts of false-positives are still detected.

4. METHODOLOGY

iWatch is a web-crawler, or spider [19], implemented in Java, and built from the ground up to search for and index data-handling practices. Similar to most crawlers, which search for and index key words, or all words within the body of a document, iWatch is designed to look for certain HTTP tokens, or HTML constructs and patterns, which may identify certain data-handling or collection techniques of interest.

Like any web-crawler, iWatch starts with a seed-list, or given set of URL's which to visit initially. iWatch downloads these pages in parallel using multiple threads, and searches the resulting download stream for web-links and a set of filters. This process is partially done using Java's built in classes and their data-handling functions (such as finding links in a HTML document), and a set of full-text searches using regular expressions.

Links found are added to a database of pages to potentially crawl. Given that most websites are complex in structure, iWatch seeks to analyze a number of pages within each domain in order to get a more complete picture of the site. At the same time, iWatch seeks to minimize the impact on the servers studied by limiting the number of pages requested from any domain. This also ensures that iWatch does not get stuck analyzing big sites, ensuring we get a minimum breadth of coverage. When a thread is idle, or is done analyzing its current page, it consults the database of links found, selecting the next eligible link and repeating the process.

Because the initial seed-list used has a tremendous effect on the overall crawling pattern it is important to choose carefully. Given the limited resources of a university/research setting, the crawler will only be able to visit a very limited number of pages and domains when compared to dedicated operations such as Google and MSN. The seed-list must therefore be selected so that the sample taken is a) as representative as possible, b) as relevant as possible, and c) leads down a path of diversity of sites.

These criteria are not always achievable. A fully representative sample would require a random sampling, which is not possible with a web-crawler, which by its nature investigates clusters of websites by following the links between these. Instead, we have chosen to construct our seed-list based on the data's potential value or impact. In other words, we ensure that the most popular sites, the sites most likely to impact the privacy of the most users, are at the heart of the crawl. In addition, to avoid an overwhelming US and English language bias, the sample must be balanced to include different countries and classes of websites. For our experiments, the crawler was seeded with a combination of the top 50 websites for that month (as determined by the Comscore MediaMetrix (http://www.comscore.com/metrix)), and a hand-picked set of popular European and Asian sites. This is far from a perfect selection of sites, but gives us an interesting and relevant sample to study.

Given a functioning web-crawler, one then needs a set of search criteria to index the pages. Table 1 gives an abbreviated list of the main bits of information we currently collect using iWatch. Many of these are composed by multiple regular expressions of mechanisms. For instance, cookies are identified by one of three filters, depending on whether they are session cookies, 1st party cookies, or 3rd party cookies. For each of these, different information is collected. iWatch collects information on 21 data-practices plus assorted site-characteristics such as geographic location based on IP address matching.

Our indices were derived from the filters used in the privacy-protecting proxy server called Privoxy (http://www.privoxy.org). Privoxy is an open-source proxy server designed to act as a filter between a browser and the web. In order to do this, Privoxy filters incoming and outgoing HTTP communication using a set of regular expressions identifying potentially dangerous or undesirable practices from an end-user perspective. These filters were manually tuned to remove some false-positives (especially in the area of webbugs) and give us more information to process.

Table 1: Main iWatch index terms

Index Terms	Description
Cookies	Identifies the use of different types of cookies (session, 1st party and 3rd party), and their characteristics
Unsolicited popups	Identifies the use of unsolicited popup windows
Webbugs	Identifies the use of third part resources potentially used to track users from site to site
Banners	Identifes the use of different types of banners and ads, potentially used to track users from site to site
P3P policies	Identifies the use of both full and compact P3P privacy policies in HTTP header
Privacy Seals	Identifies the use of Privacy seals (TRUSTe, BBBOnline, and WebTrust) in a domain's pages (link and graphic)
Data-sharing networks	A collection of the techniques used to track users across sites (3rd party cookies, webbugs, banners), and who the data is shared with
Link structure	Basic information on page's link structure and relationships between sites
Geographic information	Maps a domain/server's IP address to a country using the GeoLite database created by MaxMind (http://www.maxmind.com/)

Based on early experiments, we learned that in order to correctly identify P3P and privacy seal use, we needed to adopt a strategy other than filters. While filters effectively identify the use of compact and embedded P3P policy references, finding and downloading full P3P policies requires additional steps, which are prone to errors. As pointed out in [10], some servers will at times refuse to serve some full P3P policies from the default location (http://server/w3c/p3p.xml), skewing results. In order to ensure more correct results, we wrote a custom application that revisited each of the domains in our samples 3 times trying to get a full p3p policy. These repeated queries made a significant difference in our results, giving us an additional 117 policies for our 2006 sample, and 211 additional policies in the 2005 sample when compared with a single visit strategy. Responses were analyzed to check that what was returned was an xml document and not a html document, and that redirects were followed correctly. In the current version of the crawler, the P3P policies are not analyzed.

Our early attempts at determining seal usage directly from the pages we crawled also proved to be an ineffective strategy. Seals

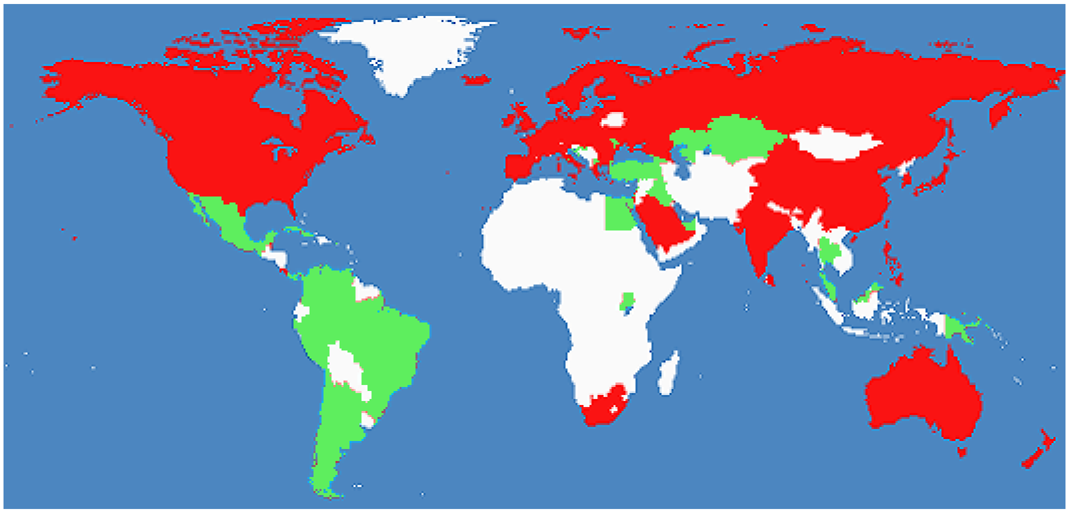


Figure 1: Geographic distribution of sample

Countries marked in red are included in the study. Countries marked in green were reached, but excluded from the study due to small sample size. Map courtesy of world66.com

are typically confined to a disclaimer or privacy policy page, therefore our ability to detect seal use through filters depends on a) the crawler having reached a policy page for the site, and b) that the seal is presented using a standard format. Of these, the first hurdle proved to be the most significant and eventually insurmountable obstacle to this strategy. To overcome these limitations we gained access to lists of certified sites directly from the certifying agencies (in this case TRUSTe and BBBOnline). These lists (http://www.truste.org/about/member_list.php, and http://www.bbbonline.org/consumer/pribrowse.asp) were then cross-referenced with our sample sites. We were unable to obtain lists for other seal providers, though this is something which we will seek to work on in the future.

To demonstrate the effectiveness and value of this approach to the study of online privacy, online regulation, and online data collection practices, we performed two experiments, one in May of 2005, and the second in August 2006, where we collected information on web-sites' privacy and data-collection practices. Each of these crawls was performed over a period of 10-14 days, with our crawler running on a single dedicated server. In this paper we will use these two samples to examine the changes that have taken place online over the last year.

Before concluding this section we wish to say a few things about the statistical testing using our two sample databases. Given the large size of our two samples, finding statistical significance is relatively simple even for relatively small changes in behavior. The reader is therefore advised that it is important to make a distinction between statistically significant and meaningful changes when considering this data. We therefore, uncharacteristically within the field of computer science, choose to set our threshold for statistical significance at the p<0.001 level throughout this paper, unless otherwise noted.

5. RESULTS

5.1 Sampling Results

Across both samples, a total of 240,340 web pages were crawled, from a total of 24,990 unique domains. There was an overlap of 1,223 domains between the two samples, or 4.7% of the total sample domains, for a total of 26,213 non-unique domains across both samples. Given that these samples were taken 15 months apart, and the speed with which websites evolve, we decided to use the non-unique total in our calculations, and treat the two samples as statistically independent. This means that on average we analyzed 9.17 pages per domain, a relatively solid basis for drawing conclusions about any given domain. Table 2 summarizes the basic characteristics of the two samples.

Overall, our two samples reached 81 countries or territories, 69 in the first sample and 60 in the second, despite the crawler being primarily seeded with US web-sites (Figure 1 shows an overview of our geographic reach). Many of these countries were represented by extremely small number of domains and pages in our data-sets, which forced us to filter some of the data to avoid drawing conclusions on overly thin data. We decided to exclude from analysis any country which was not represented by more than 10 domains across both samples, unless they were part of the European Economic Area (EEA).

Table 2: Data sample summary statistics

	Sample 1	Sample 2	Total
Collection	May 2005	August 2006	
Web-pages	119,237	121,103	240,340
Domains (unique)	15,792	10,421	26,213 (24,990)
Web-Pages/Domain (unique)	7.55	11.62	9.17 (9.62)
Total Countries	69	60	81
Filtered Countries	43	43	47
Domains/Country	367.26	242.35	557.72

The EEA is composed of the 25 European Union (EU) members, plus Iceland, Liechtenstein and Norway. All domains belonging to any EEA country were included in our sample because all EEA countries are signatories to the EU privacy directive [17], and therefore have similar privacy legislation in place. For the purpose of this analysis, the EEA countries will be viewed as a block. Of the 28 EEA countries, we found 27 in our sample (Lichtenstein being absent, see table 3 for list of all countries included in study). EEA countries make up 9.66% of our total sample.

Applying the above filtering rules, we lose 56 domains and 26 countries from Sample 1, and 34 domains and 17 countries in sample 2. Overall, 34 countries were filtered from the combined data-set, leaving 47 (43 in each of the samples). On average, the excluded countries were only represented by 2.64 domains. As could be expected, our probes primarily reached the most net-active countries in world. Though we only saw a total of 47 countries, those countries account for more than 96% of all active domains (http://www.webhosting.info/domains/country_stats). This means that though our samples only reached approximately 0.019% of all registered domains, these samples are representative of a large percentage of the net.

Table 3: Geographic distribution of sample and bias

Countries highlighted in grey to indicate EEA membership.

Based on a test of proportions, and # in the bias column together with green and tan highlight indicates significant positive or negative bias (P<0.001) respectively*

Country	Total		Samples		Bias (% of expected)
	Number of Domains	% of Domains	Number of Domains	% of Domains	
United States	46,036,912	67.56%	21,949	83.73%	* 123.94%
EEA	12,526,739	18.38%	2,531	9.66%	# 52.52%
Germany	4,039,278	5.93%	416	1.59%	# 26.77%
United Kingdom	2,947,932	4.33%	930	3.55%	# 82.01%
Canada	2,495,501	3.66%	585	2.23%	# 60.94%
China	2,099,671	3.08%	114	0.43%	# 14.11%
France	1,733,082	2.54%	197	0.75%	# 29.55%
Australia	1,393,853	2.05%	177	0.68%	# 33.01%
Spain	884,969	1.30%	210	0.80%	# 61.69%
Japan	871,196	1.28%	213	0.81%	# 63.56%
Korea	837,088	1.23%	171	0.65%	# 53.10%
Hong Kong	763,480	1.12%	27	0.10%	# 9.19%
Italy	721,992	1.06%	43	0.16%	# 15.48%
Netherlands	547,838	0.80%	157	0.60%	# 74.50%
India	342,735	0.50%	102	0.39%	77.36%
Denmark	263,789	0.39%	40	0.15%	# 39.42%
Russia	240,386	0.35%	31	0.12%	# 33.52%
Sweden	209,208	0.31%	63	0.24%	78.28%
Switzerland	186,619	0.27%	62	0.24%	86.36%
Norway	172,123	0.25%	289	1.10%	* 436.47%
Austria	163,612	0.24%	37	0.14%	58.79%
Poland	141,423	0.21%	14	0.05%	# 25.73%
Finland	123,288	0.18%	22	0.08%	# 46.39%
Belgium	122,048	0.18%	37	0.14%	78.81%
Czech Republic	91,051	0.13%	12	0.05%	# 34.26%
Israel	81,883	0.12%	39	0.15%	123.81%
Bulgaria	81,290	0.12%	2	0.01%	# 6.40%
Ireland	73,363	0.11%	21	0.08%	74.41%
Portugal	56,850	0.08%	5	0.02%	# 22.86%
New Zealand	53,517	0.08%	14	0.05%	68.00%
South Africa	48,384	0.07%	13	0.05%	69.85%
Taiwan	48,254	0.07%	34	0.13%	183.17%
Romania	35,479	0.05%	8	0.03%	58.62%
Hungary	31,249	0.05%	5	0.02%	41.59%
Saudi Arabia	29,696	0.04%	30	0.11%	262.62%
Greece	27,661	0.04%	8	0.03%	75.18%
Philippines	25,859	0.04%	17	0.06%	170.90%
Luxembourg	23,819	0.03%	5	0.02%	54.57%
Gibraltar	19,162	0.03%	2	0.01%	27.13%
Costa Rica	19,152	0.03%	16	0.06%	217.17%
Estonia	14,640	0.02%	1	0.00%	# 17.76%
Lithuania	9,988	0.01%	2	0.01%	52.05%
Slovakia	9,892	0.01%	1	0.00%	26.28%
Latvia	8,332	0.01%	1	0.00%	31.20%
Sri Lanka	5,821	0.01%	41	0.16%	* 1830.99%
Malta	5,813	0.01%	1	0.00%	44.72%
Iceland	3,047	0.00%	2	0.01%	170.63%
Sample Total	**68,142,225**	**96.34%**	**26,213**	**100%**	
Global Total	**70,733,538**				

Table 3 shows the distribution of domains across countries, as well as the bias of the sample relative to the countries current (October 2006) internet footprint. As noted earlier, the sample is skewed in favor of US web-sites, and as a consequence many other countries are underrepresented (highlighted in shades of orange in Table 3), including most EEA countries (highlighted in light grey in Table 3). Some smaller countries, through quirks of the way websites link to each other, or current events at the time of data-collection, are over-represented in the sample. As an anecdote, the bulk of our Sri Lanka sample was collected during May 2005, when peace negotiations efforts were receiving widespread press.

Given the size of the sample we collected, and the fact that there was only minimal steering of the crawler through the initial seed-list, we expected there to be significant bias in our sample when compared to the real-world. Though, as Table 3 shows, the bias in our sample is statistically significant at the p<0.001 level for approximately half of the countries in our sample (predominantly among the most net-populous nations), this bias was less than we had expected. This shows that great care needs to be taken in ensuring a seed-list which is geographically proportionate, at least for the top 20 countries (each representing more than 0.50% of the overall global domain-population). Once we exit this exclusive group, quirks and bias are less important, given the small relative size of these countries. For instance, while Norway is over-represented with 223 domains (436.47% of the sample size we should have seen), this only accounts for 0.85% of the overall sample size. This is negligible when compared with the US sample, overrepresented by 4240 domains (123.94% of the expected sample size), or 16.18% of the overall sample size. The main source of bias in our sample stems from the US being heavily over-represented. Most other countries and regions are consequently underrepresented.

5.2 Data Practices and Evolution

These data-sets have the potential to facilitate the tracking of trends in data collection practices, to gauge the effect or adoption of new technologies, new legislative requirements, best practices, and help determine if we are seeing the intended or desired effects on practices on a national or global scale. Such an analysis requires historical data going back far enough to judge long-term and short-term effects, and enough detail to determine specific causes. Our current data-set only spans 1 year, and does not, to the best of our knowledge, span any immediately obvious legislative event of relevance, making it difficult for us to perform an in-depth analysis here as proof of concept. Instead, we will focus on identifying overall trends rather than testing a specific hypothesis.

Table 4: Global data-practices

Table shows % of domains adopting practices, and the geographic spread of these practices as % of all countries in our sample.

*Based on a test of proportions a * with green highlight indicates statistically significant increase from one year ago (P<0.001)*

(1) Note that the sum of cookies used is not the same as the sum of Session, 1st, and 3rdparty cookies, as sites may set multiple cookies of different types.

	2005		2006	
Practice	**Domains**	**Countries**	**Domains**	**Countries**
Any P3P Use	24.84%	72.09%	25.90%	60.47%
Only Compact P3P Policy	1.37%	27.91%	* 1.83%	18.60%
Only Full P3P Policy	17.43%	72.09%	17.13%	58.14%
Compact & Full P3P Policy	6.05%	32.56%	* 6.94%	20.93%
Any Privacy Seal	1.99%	11.63%	* 2.03%	11.63%
Truste	0.73%	6.98%	0.95%	9.30%
BBBPrivacy	0.12%	2.33%	0.16%	2.33%
BBBReliability	0.46%	4.65%	0.92%	6.98%
Any Cookie (1)	24.03%	72.09%	* 29.08%	86.05%
Session Cookies	18.02%	72.09%	* 23.07%	86.05%
1st party Cookies	4.74%	53.49%	* 6.11%	51.16%
3rd party Cookies	3.53%	41.86%	* 5.76%	39.53%
Popups	23.59%	72.09%	24.61%	81.40%
Webbugs	33.85%	81.40%	34.52%	86.05%
Banners	8.73%	55.81%	* 10.31%	58.14%

Table 4 gives an overview of the most common and relevant data-practices with the potential to affect end-users' privacy (both negatively and positively). In this table we see both the prevalence of the data-practices for the two samples (as percentage of total domains exhibiting data-practice), as well as their geographic spread (as percentage of countries where at least 1 domain exhibits this data-practice).

Our first finding is that P3P is alive and well, with adoption among the sites in both our samples circling 25%. There were no statistically significant changes in adoption rates overall from 2005 to 2006, though the use of Compact Policies, with or without Full policies did increase significantly. These high adoption rates are likely in part due to the ubiquitous Microsoft IE 6 web-browsers' inclusion of P3P as a factor in blocking some types of cookies. Another area of good news is that though the use of compact policies is growing, use of the more expressive and meaningful Full policies dominates by a large factor.

Using our new and improved seal matching technique we see a small, but statistically significant increase in the use of privacy seals. We realize that our list of seal providers is simplistic and short, and that more providers need to be added in order to provide a more realistic picture of the use of seals today. As a point of contrast, others [1] have found that 11% of US websites had privacy seals in 2001. It is unlikely that seal adoption has decreased this significantly over the last 5 years.

Looking at the much maligned cookie, we see that overall use has increased markedly over the course of the year. This increase is seen both in the use of inoffensive session cookies as well as the more troubling 3rd-party cookie. We also see more sites using more than one type of cookie, though we have not computed statistics on how many cookies of the same type a site uses. The one bright note to raise here is that though the number of domains using 3rd party cookies grew, geographic distribution declined.

As expected from the improvements seen in terms of online ad revenues in the past year, we see a significant growth in the number of domains using banner ads. On the other hand, the use of unsolicited popups and webbugs is flat from a year ago, though geographic distribution is up.

Table 5: Effects of P3P and Privacy Seals on practices

Table shows % of domains adopting practices, the expected rates (product of the probability of the two practices), and the difference (diff) from this expected rate.

*Based on a test of proportion, cells marked by *or # with green or tan highlight in 2006 "Detect" column indicates statistically significant increase or decrease from one year ago (p<0.001, 2-tailed)*

Based on Chi-Square tests of independence, combinations marked with a ^ and highlighted blue in the "diff" columns were not statistically independent (P<0.001)

	2005			2006		
Practices	**Detect**	**Expect**	**diff**	**Detect**	**Expect**	**diff**
P3P+Webbugs	11.99%	8.41%	^ 142.6%	* 13.75%	8.94%	^ 153.8%
Seal+Webbugs	0.96%	0.44%	^ 217.0%	0.92%	0.32%	^ 289.7%
P3P+Popups	11.61%	5.90%	^ 196.9%	12.15%	6.37%	^ 190.6%
Seal+Popups	0.89%	0.31%	^ 286.9%	1.13%	0.50%	^ 226.1%
P3P+Session C	4.51%	4.48%	100.8%	* 5.70%	5.97%	95.4%
Seal+Session C	0.41%	0.24%	^ 174.2%	* 0.86%	0.47%	^ 184.0%
P3P+1st party C	1.48%	1.18%	^ 125.9%	1.66%	1.58%	104.9%
Seal+1st party C	0.24%	0.06%	^ 387.6%	0.33%	0.12%	^ 262.4%
P3P+3rd party C	1.61%	0.88%	^ 183.2%	* 3.22%	1.49%	^ 216.2%
Seal+3rd party C	0.24%	0.06%	^ 228.3%	* 0.51%	0.12%	^ 434.2%
Seal+P3P	0.60%	0.33%	^ 184.7%	# 0.33%	0.53%	^ 61.9%

Some of the most interesting findings from our study deal with the effect that the use of P3P and privacy seals has on the prevalence of other data-practices. What we are looking for here is whether the group of other practices is statistically independent from the use of privacy seals or P3P policies. In Table 5 we present the basic data, as well as the results of tests of proportions seeing whether the rate increased or decreased from one year to the other, and Chi-Square (test of independence) to determine whether the differences between observed or detected rates and expected rates differ in a statistically significant way.

As Table 5 shows, P3P and privacy seal use was not statistically independent from most of the other privacy indicators examined in this study. The presence of either of these indicators was usually associated with a positive co-occurrence rate. This may have had (and likely does have) a perfectly reasonable explanation in that sites with more complex information needs and data collection practices seek to assure and explain the use of other technologies through a P3P policy, or provide assurance of their intent through the presence of a seal. Because P3P policies were not analyzed in this study, we cannot say whether policies addressed or explained the use of the correlated technologies, though this is something which should be investigated in the future.

From 2005 to 2006 we saw a statistically significant increase in the use of P3P in conjunction with webbugs, session cookies, and 3rd party cookies, while the same was observed for privacy seals and session cookies and 3rd party cookies. This represents a mixed bag for end-users, as both desirable and undesirable practices showed an increase. On the other hand, the co-occurrence of privacy seals and p3p policies decreased significantly from 2005 to 2006, part of an observed trend in avoiding overlapping certification or explanation systems.

The prevalence of P3P use was an issue which we decided to explore in greater depth. Specifically, we wanted to explore to what extent P3P use was constrained, or influenced by the site's popularity (as defined by our seed-list selection). By partitioning the domains crawled into segments of 1000 domains we get a rough ranking of the sites (see Figure 2). This is dependent on the acceptance of a definition of popularity being the distance from the seed-list sites. While not a fully fair metric, it does fit with the way browsing patterns affect page rankings, and is probably good enough for the purposes of this investigation. As can be seen in Figure 2, popularity does indeed affect the adoption of P3P, though much more markedly today than in 2005.

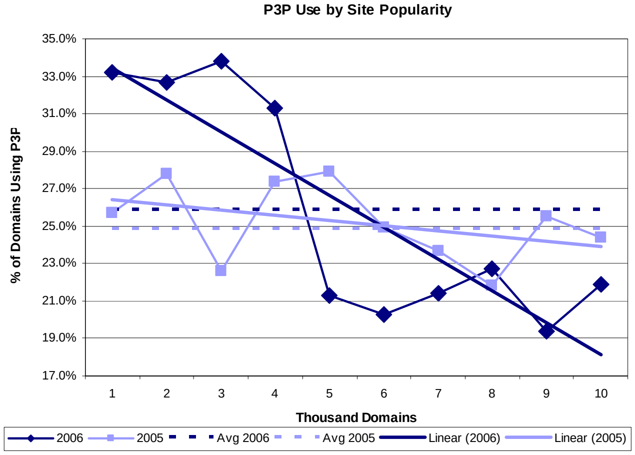


Figure 2: P3P use by site popularity

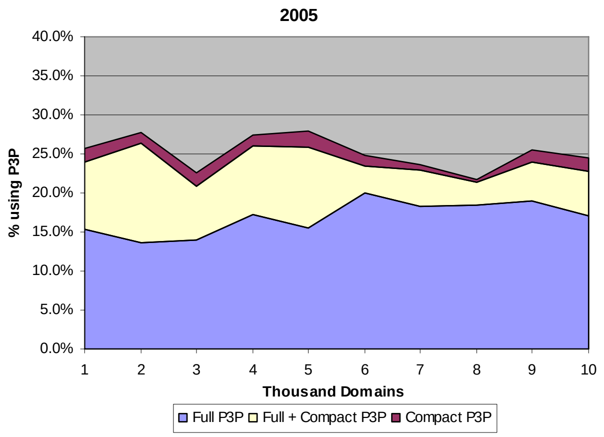


Figure 3: P3P use by site popularity and type, 2005

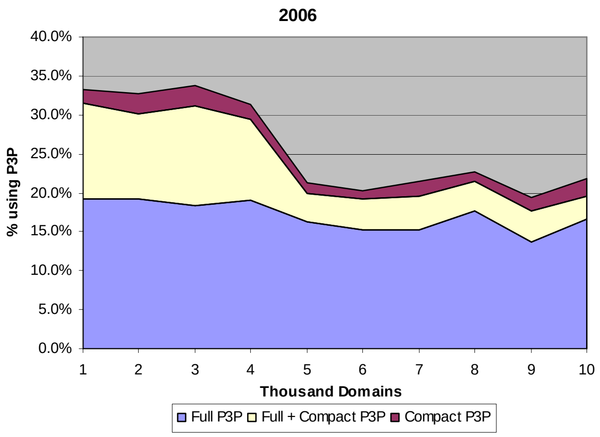


Figure 4: P3P use by site popularity and type, 2006

Figures 3 and 4 show how the use of P3P has evolved from 2005 to 2006 in terms of the types of P3P policies used, and the popularity of the sites using them. From figure 3 we can see that in 2005 as a sites' popularity decreases, fewer offer dual policies (fewer sites offer compact policies), instead offering only full policies. From figure 4 we can see that the increase in P3P use observed over the two samples is in large part due to a significant increase in the pre 4,000 sites, which are offering more dual and full policies. Beyond this, the distributions look very similar.

5.3 Legislation and Data Practices

As previously mentioned, one of the intended uses of these data-sets is to examine the effects that legislation and regulation have on data-practices. Given that no major new US privacy legislation took effect between our two samples, we instead use our samples to examine the privacy practices, and evolution of these between the US, Canada, the UK, and the EEA, all countries or regions with different levels of legislation regulating data-practices and the collection and use of PII. Table 6 gives an overview of the geographic clustering of data.

The most interesting elements for this analysis is the EEA and US columns, as they represent two ends of the spectrum in terms of privacy regulation and enforcement activity. The UK and Canadian samples are interesting because they serve as interesting points along this continuum. Both the UK and Canadian privacy regulations are stricter than those seen in the US, yet both are influenced by similar culture, language, technology adoption, etc. If legislation and user activism have an effect on the adoption of technologies and practices, we should see some systematic differences in this data, especially between the US and EEA.

Table 6: Geographic clustering of domains

Table shows number of countries and the % of all domains in each group and sample. In the total column we give the actual number of domains.

** UK appears both on its own and as part of the EEA sample*

*Based on a test of proportion, cells marked by *or # with green or tan highlight in 2006 Detected column indicates statistically significant increase or decrease from one year ago (p<0.01)*

	2005		2006		Total	
Geographic Area	Country Count	Domains	Country Count	Domains	Country Count	Domains (unique)
EEA	24	9.75%	25	9.52%	27	2,531 (2,483)
Canada	1	2.41%	1	# 1.96%	1	585 (576)
United Kingdom*	1	3.18%	1	* 4.11%	1	930 (899)
United States	1	83.28%	1	* 84.43%	1	21,949 (20,815)
Other	17	4.57%	16	4.10%	17	1,148 (1,117)

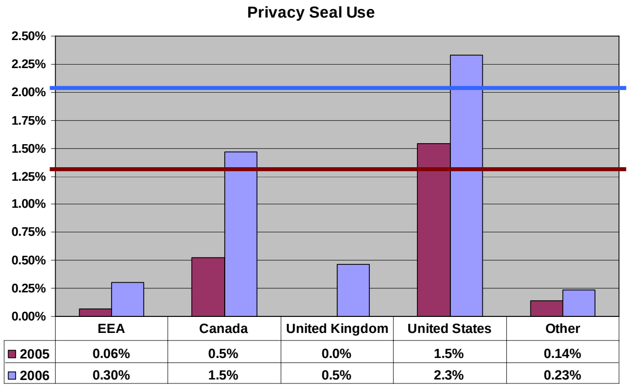


Figure 5: Privacy seals by geographic area

Horizontal bars showing global average for the two samples (by color). All changes from 2005 to 2006 except 'Other' category are statistically significant (p<0.005)

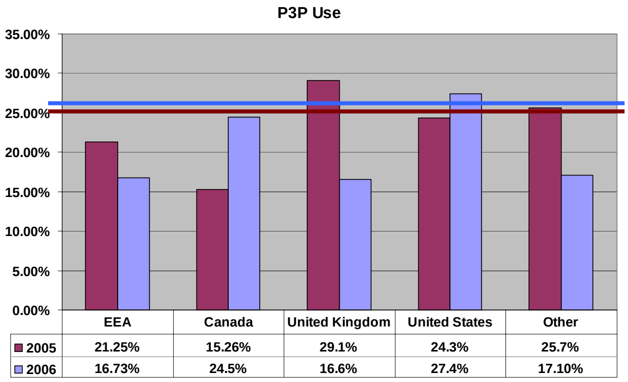
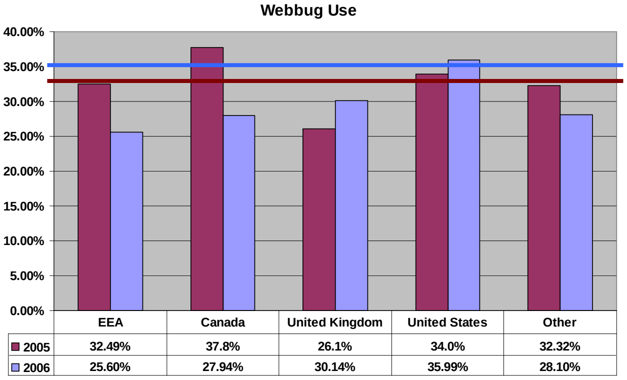


Figure 6: P3P adoption by geographic area

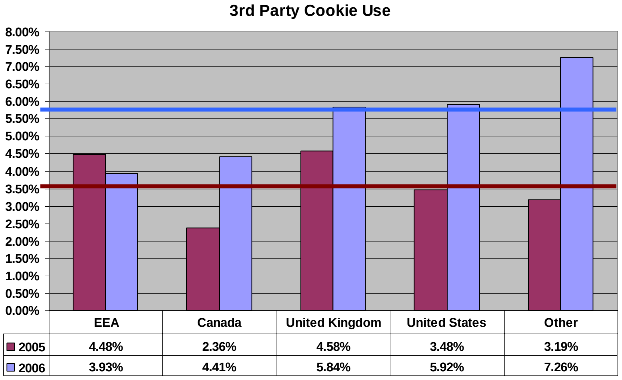
Horizontal bars showing global average for the two samples (by color). All changes from 2005 to 2006 are statistically significant (p<0.005)



	EEA	Canada	United Kingdom	United States	Other
🟥 2005	32.49%	37.8%	26.1%	34.0%	32.32%
🟦 2006	25.60%	27.94%	30.14%	35.99%	28.10%

Figure 7: Webbug use by geographic area

Horizontal bars showing global average for the two samples (by color). All changes from 2005 to 2006 are statistically significant ($p<0.005$)



	EEA	Canada	United Kingdom	United States	Other
🟥 2005	4.48%	2.36%	4.58%	3.48%	3.19%
🟦 2006	3.93%	4.41%	5.84%	5.92%	7.26%

Figure 8: 3rd party cookie use by geographic area

Horizontal bars showing global average for the two samples (by color). All changes from 2005 to 2006 are statistically significant ($p<0.005$) except for EEA group ($p<0.05$).

Some of the interesting observations are that, as Figure 5 shows, privacy seals are virtually non-existent outside of the US and Canada. Again, data for the use and adoption of privacy seals is incomplete and should be viewed with caution, but we would expect these deficiencies to play out evenly geographically, as all major certification agencies are US based. It is interesting to note that the only countries to use privacy seals in 2006 were the US, UK, South Africa, Canada and Belgium. Apart from the later, these are all countries where English is (one of) the official languages. In 2005, privacy seal use was restricted to the US, Canada, Japan, and Finland.

Another interesting finding is the skew in P3P adoption, with the US and Canada very much leading the way (Figure 6), with every other region showing a statistically significant decline. Determining why this is the case could be an interesting issue to investigate in the future, and would also require the analysis of the P3P policies themselves.

While other technologies could have been examined in this fashion, we decided to conclude this study by looking at two technologies which are particularly problematic for end-user privacy; webbug and 3rd party cookie use. Again, if regulation affects web-based data practices, this is where we should expect to see the biggest differences (see Figures 7 and 8). While the observed trends were in line with our expectations, the differences were not as marked as we had expected, nor were they uniform. The UK, a part of the EEA sample, consistently followed the patterns exhibited by the US rather than its European partners.

As noted at the beginning of this section, the impact of legislation on these practices remains a question which warrants further investigation. The short time spanned between the samples, the fact that at this point there are only 2 samples, and that no major piece of legislation was enacted which directly impacted online privacy practices, made it difficult for us to explore this use fo the data. With time however, we believe it will be interesting to investigate the long-term effect of legislation such as the GLBA on financial sites, or HIPAA on healthcare sites. This will however require a more longitudinal sampling method (given that both laws were in force when our first sample was taken), and a stronger focus on financial and healthcare sites.

6. DISCUSSION

The goals of this paper were to demonstrate the feasibility and value of using a system such as iWatch to study the current state of the art in terms of online practices and data collection techniques which may affect end-user privacy, and to provide a minimum set of current data about prevalent data practices. We believe we have demonstrated that the general approach is sound, though some fine-tuning is necessary. We have also generated a broad set of statistics which others may build on in their own research or system design. Having said this, a number of important lessons were learned as part of this study.

Given that we are using a web-crawler, following links as they appear on web-pages, our sample of domains is always going to be different from one crawl to the next. It is therefore difficult if not impossible to precisely control the distribution of sites. This presents two potential problems. The first is that it is difficult if not impossible to get a completely unbiased sample (at least in terms of geographic representation) by chance. Though for our purpose, some small adjustments are likely to be enough; those with a need for greater accuracy can enforce the distribution they desire by sampling from the dataset to achieve the right proportions of sites, though this would reduce the size of the overall dataset.

The second potential problem is that because of the dynamic nature of the web, any two samples are likely to deviate significantly in terms of the sites visited. If this deviation takes place early enough in the process, it may be difficult to directly compare samples. As an example, imagine that a significant number of the seed-list sites in instance A link to academic sites (due to some ongoing news story). In instance B, the same seed-list may instead point to a collection of e-commerce sites instead. In our samples, we had a seed-list of 100 items each time. Half that seed-list came from a public top-50 site list, and half the sites were manually picked to ensure a greater geographic distribution. Even though these samples were only separated by a year, there was only an 36% overlap in the top-50 site portion of the list. This likely lead to a significant divergence of the two samples, and possibly false inferences about changing practices, if the sample site is too small. With a large enough sample size, all things should even out.

This brings us to the question of whether a sample size of 0.02% of all domains is adequate for this kind of analysis. As a proof of concept we were more than happy with this sample size, though for a production and archival system that may not be sufficient. While efforts to streamline data-collection, and thereby the resulting sample size can and will be made, the question of how much data must be collected and will need to be revisited.

One important area of bias which is not represented in Table 3, and for which we have no measure, but may nevertheless be of concern, is the likely under-representation of different market segments and domain types. Our seed-list was composed of the most popular websites of the day, all belonging to major corporations. Smaller "mom and pop" or non-commercial sites are therefore likely underrepresented. Previous research has shown that the web is not a completely connected graph. Rather, the web is a set of disconnected islands [18]. We therefore depend on a well-chosen seed-list to ensure that we can reach as many of these islands as possible, and have to accept that some sites will never be reachable. This is a possibility which concerns us, though the most popular websites are probably most important to most, a balanced, diverse sample would be more valuable overall.

We are also concerned about the difficulties we experienced in collecting full P3P policies, and the errors this could introduce into the analysis. We found that by trying to access full policies 3 times we got a significantly larger number of policies, but how many times should we try and access a server before giving up? Would we have found even more policies if we had checked back 5 times, 10, or 100? This instability is a problem which the community will have to address if P3P is to see further gains in adoption.

While there has been much debate about the value and shortcomings of P3P, the authors' perspective is that the adoption of technologies which communicate potential problems to the end-user (even if as some argue, flawed) can only be a positive thing. We were especially intrigued to find that the use of P3P policies coincided with the use of other, less desirable data collection practices such as 3rd party cookies and webbugs. Determining what the role of the policy was in that relation (smokescreen or explanation mechanism) is an interesting open question, one that would require us to parse the P3P policies.

Our inability to parse the P3P messages and compare their content to observed practices in time for this study is a significant shortcoming, and one which we will address in future work. Without knowing what P3P policies actually specify, and whether they contradict actual practices we cannot draw any solid conclusions as to the correlation between P3P adoption and things like 3rd-party cookies and webbugs.

We were reasonably pleased with our success with identifying sites using privacy seals (using official published lists from certifying agency). Early experiments trying to detect seals in the HTML stream yielded only a fraction of the sites found by matching against the seal providers lists, at a fraction of the cost. On the down-side side, our numbers are much lower than those reported by some others, leading us to conclude that in order for this to be a viable approach we need to broaden our list of seals. Search for seals in the HTML was appealing from the perspective of looking for misuse of seals, but this in retrospect turned out to be too difficult to do automatically. In [29], the reported detection of unauthorized seal use was performed manually, an approach which does not work with our intent of large-scale analysis. Automatically analyzing images unambiguously is very difficult, leading us to abandon these efforts.

Despite these shortcomings, our analysis also showed that there are interesting trends and patterns worth investigating with these datasets. One of the areas which we hope to expand into is the identification of best practices and guidelines to developers, legislators, and users. We also believe that these datasets could be of use to developers of privacy protection tools to either provide training or seed-date for more intelligent recommendation systems, or to inform where efforts are best spent.

One potential shortcoming to this geographic analysis is that our server is based in the US. In cases where we crawl multinational corporations or mirrored sites, our crawler is going to get directed to a US-based mirror. Given that we use GeoIP (www.maxmind.com) to map IP addresses to geographic locations for the servers, our results are necessarily be somewhat skewed, especially given that the sites most likely to engage in such behavior are the sites in our seed-list. Unfortunately, we do not at this time have a remedy for this problem.

7. FUTURE WORK

We have throughout this paper identified a number of shortcomings and caveats to our approach, discussing these where it seemed most natural. Our goals for the coming months are therefore relatively clear. We believe this study has validated the general approach, though some of the implementation details need to be refined. Our goal therefore is to address these as soon as possible so we can start to offer these data-sets to researchers, policy makers and tool developers on a regular basis (quarterly).

One of the areas of improvement identified in this study is the need for more careful balancing of the seed-list. We believe to have a strategy which will ensure a more balanced crawl, but acknowledge the fact that to a certain extent we are at the mercy of the tides. An intriguing possibility is to force the crawler to enforce the geographic proportions, but this would only work to ensure we do not over-represent any country or territory. There is however little we can do to ensure a minimum set of domains in a region short of stacking the seed-list.

In this study we also set an arbitrary cut-off point for countries (members of EEA, or the sighting of 10 unique domains in our sample). We now believe this policy may be less than desirable, and that instead a more reasonable policy would be to set a target for the number of domains to crawl, and close off countries or regions once their allotted quota of sites is reached. The list of links to crawl can quite easily be instrumented to keep track of the links' country of origin, which may then be used in the selection criteria.

We also need to reach out to more seal providers. While we have a short list of additional providers to contact, one difficult question is going to be again, when we have a complete enough set of seals, as well as ensuring that our list of seal certified sites remains up to date.

We believe that what we have been able to show in this paper is only the beginning of the kind of analysis which is possible with these types of data. The next steps includes looking at this data with more advanced statistical tools such as cluster analysis to look for patterns, either geographic or in terms of industries. We hope this kind of analysis can identify things like best practices, or industry conventions. This will helpfully help address some of the most serious points of criticism to this work, which is that though some of our analysis provides interesting insights, most of the data is rater superficial.

Working along these same lines we are currently trying to apply machine learning techniques to the datasets we have available, in combination with things like Netcraft's index of known phishing and malware sites, and their geographic locations. Using this data we hope to determine what meaningful risk indicators may be, in the hope of providing end-users with risk estimates before they follow a link or access a given site.

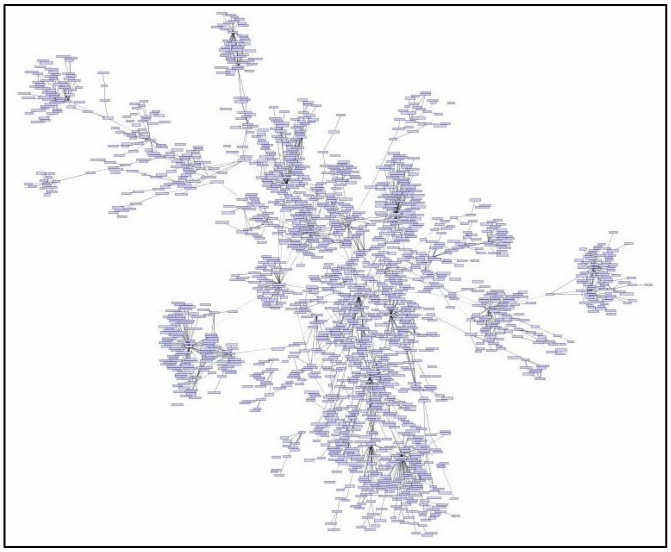


Figure 9: Data-sharing network based on cookies

Along these lines we are performing more sophisticated types of analysis, such as detecting and tracking information sharing networks composed of cookies, webbugs, banners and similar technologies. Initial explorations are promising. Figure 9 shows a network, visualized from real data, of sites connected to each other through 3rd party cookies. Each of the blue rectangles represents a domain, and each line a cookie. We were surprised by the number and size of the networks detected, and believe this can be a useful way of examining the spread of information.

We are also interested in looking for policy pages in order to try and combine our data with goals extracted from natural language policies, forming pseudo-machine readable policies, as explored by [3, 25]. These could then be compared to observed practices, and P3P policies to try and detect inconsistencies. Detecting inconsistencies between stated policy and observed practices will probably be one of the most valuable pieces of data in terms of identifying sites which put end-users' privacy at risk, either through malice or negligence.

Finally, we are naturally seeking to receive and incorporate feedback from other researchers on what practices to track and ways to track or improve the accuracy and value of our data. This could also potentially extend to accepting recommendations on new practices or technologies to track.

8. ACKNOWLEDGMENTS

This work was in part funded by NSF ITR Grant #0113792. We thank Yi Han Bae, John O. Ndukuba, Robert Marinski, and Leandro Taberner for their help in developing the iWatch crawler, as well as the support staff at The Georgia Institute of Technology and Oregon State University for their assistance. We also thank the students, colleagues and reviewers who have helped us with revisions of this paper.

9. REFERENCES

[1] Adkinson, W.F., Eisenach, J.A., and Lenard T.M. *Privacy Online: A Report on the Information Practices and Policies of Commercial Web Sites*. Progress and Freedom Foundation, Washington DC. March 2002.

[2] Anderson, R.E. "Social impacts of computing: Codes of professional ethics." *Social Science Computing Review*, 2 (Winter 1992), 453-469.

[3] Antón, A.I., Earp, J.B., Bolchini, D., He, Q., Jensen, C., and Stufflebeam, W. "The Lack of Clarity in Financial Privacy Policies and the Need for Standardization." *IEEE Security & Privacy*, 2(2), pp. 36-45, 2004.

[4] Arshad, F. "Privacy Fox - A JavaScript-based P3P Agent for Mozilla Firefox." *Privacy Policy, Law, and Technology.* 17-801

[5] Ashley, P., and Schunter, M. "The Platform for Enterprise Privacy Practices." *Information Security Solutions Europe*, Paris France, October 2002.

[6] Belanger, F., Hiller1, J.S., Smith, W.J. "Trustworthiness in electronic commerce: the role of privacy, security, and site attributes." *Journal of Strategic Information Systems* 11 (2002) 245–270.

[7] Campbell, A.J. "Relationship marketing in consumer markets: A comparison of managerial and consumer attitudes about information privacy." *Journal of Direct Marketing* 11, 3 (Summer 1997), 44–56.

[8] Cranor, L.F, Langheinrich, M., Marchiori, M., Presler-Marshall, M., and Reagle, J. The Platform for Privacy Preferences 1.0 (P3P1.0) Specification. Retrieved Nov 10, 2004. http://www.w3.org/TR/P3P.

[9] Cranor, L.F. *Web Privacy with P3P*. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 2002.

[10] Cranor, L.F., Bayers,S., Kormann,D. "Automated Analysis of P3P-Enabled Web sites" Proceedings of the 5th *International Conference on Electronic Commerce, ICEC2003*

[11] Culnan, M.J. "Georgetown Internet Privacy Policy Survey: Report to the Federal Trade Commission." Washington, DC: Georgetown University, McDonough School of Business.

[12] Culnan, M. J. and Milne, G. R. "The Culnan-Milne Survey on Consumers & Online Privacy Notices: Summary of Responses." Washington DC: FTC, December 2001.

[13] Dhamija, R.,and Tygar, J.D., ."The battle against phishing: Dynamic Security Skins." Proceedings of the *2005 Symposium on Usable Privacy and Security*.

[14] Dhamija, R., Tygar, J.D., and Hearst, M. "Why Phishing Works." In *Proceedings of CHI 2006*, April 22–27, 2006, Montréal, Québec, Canada.

[15] Earp J.B. and Meyer, G. "Internet Consumer Behavior: Privacy and its Impact on Internet Policy", *28th Telecommunications Policy Research Conference*, Sept. 23-25, 2000.

[16] Egelman, S., Cranor, L.F., and Chowdhury, A. "An Analysis of P3PEnabled Web Sites among Top 20 Search Results." *ICEC'06*, August 14—16, 2006, Fredericton, Canada.

[17] European Union (EU). *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.*

[18] Flake, W.G., Pennock, D.M., and Fain, D.C. The Self-Organized Web: The Yin to the Semantic Web's Yang IEEE Intelligent Systems, 2003.

[19] Heydon, A., and Najork, M. "Mercator: A scalable, extensible Web crawler." *World Wide Web* Volume 2, Number 4, December, 1999.

[20] Javelin Strategy & Research, *2005 Identity Fraud Survey Report*, January 2005. http://www.javelinstrategy.com/reports/2005IdentityFraudSurveyReport.html.

[21] Jensen, C., Potts, C., and Jensen, C. "Privacy practices of Internet users: Self-report versus observed behavior." *International Journal of Human-Computer Studie*s Volume 63, Issues 1-2, July 2005, 203-227.

[22] Jensen. C. and Potts, C. "Privacy Policies as Decision-Making Tools: A Usability Evaluation of Online Privacy Notices" *Proceedings of CHI'04* Vienna, Austria, April 2004

[23] Jupiter Research. "Security and Privacy Data." Presentation to the *FTC Security Workshop*, May 20, 2002

[24] Juvenal. *The Sixteen Satires*, Satire VI, verse 347. Penguin Classics; 3rd edition 1999.

[25] Karat,J., Karat,C.M.,and Brodie,C.A. "An empirical study of natural language parsing of privacy policy rules using the SPARCLE policy workbench" Proceedings of the second *Symposium on Usable Privacy and Security, SOUPS*

[26] Kuner., C. *European Data Privacy Law and Online Business.* Oxford University Press., 2003.

[27] Meinert, D.B., and Peterson, D.K. "Would Regulation of Web Site Privacy Policy Statements Increase Consumer Trust?" *Information Science Journal* Volume 9, 2006

[28] Miyazaki, A. D., Krishnamurthy, S. "Internet Seals of Approval: Effects on Online Privacy Policies and Consumer Perceptions" *Journal of Consumer Affairs*, Volume 36 Issue 1 Page 28-49, 2002.

[29] Moores, T.T., and Dhillon, G. "Do Privacy Seals in E-Commerce Really Work?" *Communications Of The ACM* December 2003/Vol. 46

[30] Moshchuk, A., Bragin, T., Gribble, S.D., Levy, H.M. "A Crawler-based Study of Spyware on the Web" in *Proceedings of the Annual Network and Distributed System Security Symposium.* San Diego, February 2007.

[31] Plato. *The Republic*. Penguin Classics; 2nd edition 2003

[32] Provos, N., McNamee, D., Mavrommatis, P., Wang, K., Modadugu, N. "The Ghost In The Browser Analysis of Web-based Malware." *First Workshop on Hot Topics in Understanding Botnets* April 10, 2007, Cambridge, MA.

[33] Schwartz, P.M., and Reidenberg, J.R. *Data Privacy Law: A Study of United States Data Protection.* Michie, 1996.

[34] United States (US) *Children's Online Privacy Protection Act of 1998*, Public Law No. 105-277, October 21, 1998.

[35] United States (US) *Gramm-Leach-Bliley Financial Modernization Act of 1999*, Public Law No. 106-102, November 1, 1999.

[36] United States (US) *Health Insurance Portability and Accountability Act of 1996*, Public Law No. 104-191, August 21, 1996.

[37] Whitten, A. and Tygar, J.D. "Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0," *Proceedings of the 8th USENIX Security Symposium*.

[38] Weirich D. and Sasse, M.A. (2001) "Pretty Good Persuasion: A first step towards effective password security for the Real World." *Proceedings of the New Security Paradigms Workshop 2001* (Sept. 10-13, Cloudcroft, NM), pp. 137-143. ACM Press.

[39] Wu, M., Miller, R.C., Garfinkel, S.L. "Do Security Toolbars Actually Prevent Phishing Attacks?" in *proceedings of CHI 2006*, April 22-27, 2006, Montréal, Québec, Canada.