AN ABSTRACT OF THE THESIS OF

Chandan Sarkar for the degree of Master of Science in Computer Science of presented on October 26th 2007.

Title: An Automated Web Crawl Methodology to Analyze the Online Privacy Landscape.

Abstract approved: _____

Dr. Carlos Jensen

Protecting end-users privacy and building trust are the two most important factors needed to support the growth of ecommerce. The increased dependence on the Internet for a wide variety of daily transactions causes a corresponding loss in privacy for many users, as virtually all websites collect data from users directly or indirectly while performing business with them. In this thesis I have used a web crawler named "iWatch" which serves as an instrument to collect basic statistics on the state of privacy, security, and data-collection practices on the web. I have looked at several interesting practices, and ways of examining the data. This thesis is also meant to serve as a point for reflection and discussion about which practices to observe, and how the raw data from such a system can and should be evolved and made available to a wider audience. The purpose of this thesis is to show web-crawling is a valid approach to mass data collection over the internet with the aim of predicting privacy practices and analyzing how they have evolved in the last three years in terms of geography, legislation, risks, biases and flows.

Finally I demonstrate methods to show how to control bias while collecting data, and I propose a probabilistic mathematical model to limit the depth of search to achieve wider breadth for web crawling techniques in the future.

.

An Automated Web Crawl Methodology to Analyze the Online Privacy Landscape

by
Chandan Sarkar

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented October 26,2007
Commencement June 2008

Master of Science thesis of Chandan Sarkar presented on October 26, 2007

APPROVED:

Major Professor, representing Computer Science

Director of the School of Electrical Engineering and Computer Science

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Chandan Sarkar, Author

ACKNOWLEDGEMENTS

This M.S dissertation is the culmination of two years work, which would not have been possible without the help and support of my beloved parents, my uncle as well as friends, and colleagues. I would also like to thank my sister and my brother in law for their support.

I would like to thank my advisor, Dr. Carlos Jensen for being all his support and guidance. Carlos has been an exceptional inspiration for me for the last two years in this privacy research and in the field of HCI. Without his support and guidance I won't be able to reach this final destination, where I am today. Thank you once again.

I would also like to thank Dr. Margaret Burnett not just for serving as a committee member in my thesis defense, but also for her great support and guidance for the last two years and always answering to my questions with smiling face.

I would also thank Dr. Budd and Dr. Higdon for serving in my committee as well as their cooperation in all respect. I would also like to thank Dr. Wong Li from the department of Statistics and Dr. Wong and Dr. Bella Bose for their guidance and support and helping me understand many complex ideas.

Finally I would like to thank all the members of HCI research group and some of my colleagues and friends in computer science specially Rob Hess, Erin Fitzhenry, Daman Oberoi, Marie-Anne Midy, Chris Chambers, Tim Bauer for their support and brainstorming many crazy ideas with me and helping me graduate.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

An Automated Web Crawl Methodology to Analyze the Online Privacy Landscape

1 Introduction

Scholars and researchers have claimed that consumers' continuous inclination towards a consumer centric service economy with the growth of ecommerce, has not affected their views towards privacy protection [Cranor et al. 2007, Westin et. al. 2001]. Consumers' privacy concerns are mainly centered on intrusions, manipulation, third parties capturing the sensitive personal information on the internet, vulnerability to third parties who have access to their personal information, and identity theft. The driving factor behind consumers' privacy concern stem from high levels of distrust towards internet institution's data collection methodologies, their dubious data retention policies and fears of technology abuse.

Data protection is broadly analogous to the concept of information privacy which is the claim of individuals, groups or institutions to determine for themselves when, how and to what extent information about them is communicated to others [Bennett et. al 1992]. According to a 2007 U.S Senate report, 36 million identity thefts took place in the US in 2003 and 155 million personal records have been compromised since 2005. These include 600 publicly reported data breaches out of which 400 were resulted in the exposure of Social Security Numbers. [Anton et al. 2007].

These reports to some extent justify why consumers today are sensitive to privacy issues when conducting business online, despite the fact identity theft takes place offline more compared to online [Javelin Report 2005]. Trust is an important factor for the growth of ecommerce. Thus protection of information by enforcing security and privacy practices is a way for organizations to increase business by building trust with consumes [Ponnurangam and Cranor 2007]. Despite these efforts, surveys and studies have indicated that users are increasingly concerned that about their privacy when they go online [Culnan et al. 2001]. While most people claim to be very concerned about

their privacy, they do not consistently take action to protect it. Web retailers detail their information practices in their privacy policies, but most of the time this information remains invisible to consumers [Egelman and Cranor 2007].

Virtually all websites collect data from users directly or indirectly. Thus the increased dependence on the Internet for a wide variety of daily transactions causes corresponding loss in privacy for most users. Web, a global system, crossing many of the traditional lines of jurisdictions is a complex place in terms of technology and practices. As technology and business practices are constantly evolving, keeping up with changes and trends sometime seem like a mammoth task. 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. The picture gets more complicated when we consider business practices about multi-national companies, and potential business-to-business (b2b) partners. The major questions are: how are corporations handling personal information? What policies and practices are governing such information? Corporations routinely handle personal information (medical, financial, purchase records) and many times operate without policies in various areas. Also policies within the organization exist, but sometimes conflict with practices in the organization [Smith J. et. al 1993]. Business practices and technologies are constantly evolving, making it difficult for the consumer to make a judgment whom they should trust with their data.

The primary purpose of data protection laws both in Europe and United States asserts that data should be collected by lawful means and with the knowledge or consent of the individual concerned. Data should be relevant for the purpose for which they are used and these purposes should be explicitly mentioned at the time of collection, and organizations have a responsibility to maintain a reasonable security safeguard [Bennett et. al 1992]. The implication of this, the organization must ensure that the existence and nature of record keeping system are public knowledge and that data subjects can obtain

and correct any information pertaining to them which is not timely, accurate, and complete. The true caveat of privacy principles in terms data collections and protection, and retention for the user and law is a complex process.

Based on the general statement of the privacy protection it is evident that the issue of jurisdiction has been, and will continue to be for the foreseeable future, a serious challenge to the policy enforcement in e-commerce. Determining compliance should therefore be a major concern for designers, developers, and administrators of such systems. It is important for the policy maker as well as the legislator to understand the impact of privacy and practices on common mass and craft rules and legislation that are effective and meaningful. As legislation often lags behind technological adoption and development, it is important to monitor what and 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.

We 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 [Culnan et.al 2001]. The important factor is 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. Given these facts research has shown that consumers are willing to pay more if their privacy being protected [Cranor et. al 2007].

Hence the challenging factor for the researcher is 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 decisions.

Similarly for developer and system administrator it is important to understand what system and implementation is worth design and develop of, where to invest their time and money in order to escape who stand to loose significant time and money, or potentially face a user backlash and/or fines from flawed models and designs.

In order to meet the information needs of such diverse stakeholders it is absolutely essential to have a reliable set of data about current data practices and technology use. As this data is likely to influence public policy, consumer perception, as well as business practices, it is essential that the data should 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 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 [Cranor et.al 2003], though not on the scale of what we demonstrate in this thesis. The aim of this thesis is to show what are effective and efficient strategies need to be taken to collect large amount of data set to analyze the trends and evolution of internet, what are the precautions we need to take to eliminate possible sources of bias, and what kind of design and probabilistic mathematical model we should look for an automated analysis.

We designed a web crawler named iWatch in order to handle all the above mentioned problems. The name “iWatch” is derived from the famous question “Quis custodiet

ipsos custodes?” or "Who watches the watchers/guards?" originally posed by Plato in The Republicand popularized in Latin in Juvenal’s Satires [Juvenal Satire 24]. Hence, iWatch monitors those who normally monitor us; websites. iWatch serve as a source of basic statistics on the state of privacy, security, and data-collection practices on the web. As 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 thesis 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 have looked at several interesting practices, and ways of examining the data. The purpose of this thesis is to show web-crawling is a valid approach of large set of data collection over the internet to predict the privacy and security associated with it, what are the geographic trends of privacy practices internet is evolving in last three consecutive years in terms of geography and legislations, what are the risks, biases and flows associated with it and what probable measures we can take to reduce the biases.

2 Literature Review

Over the past five years researchers have begun using the automated process of web crawling to gather data from the internet. This methodology is used to analyze the evolving nature and trends of the web. Hence researchers are more and more inclined to use this web crawling technology for sampling the patterns of the internet.

2.1 Detection of Malware

In order to determine the presence of malware on the internet, researchers from Google used a web crawling approach. Using this technique they conducted a study and offered some statistics regarding the presence of web based malware on the internet. The results from the Google crawled web repository were evaluated over a period of twelve months. The results showed several attack strategies for turning web pages into malware infection vectors. Also, four main aspects of content control, which are mainly responsible for enabling browser exploitation: advertising, third party widgets, user contributed content and web server security. Through analysis and examples, they showed how each of these categories can be used to exploit web browsers [Provos et. al 2007].

2.2 Detection of Spyware

In a similar effort in 2005, researchers from the University of Washington analyzed the presence as a threat of malicious spyware on the internet using a web crawler. In order to determine how spyware had penetrated different regions of the web, their designed crawler crawled sites from eight different genres: adult entertainment sites, celebrity-oriented sites, games-oriented sites, kids' sites, music sites, online news sites, pirate sites, and screensaver or "wallpaper" sites. In addition, they also crawled c|net's download.com shareware site. These sites were selected either using Google directory or using the results of category-specific Google keyword searches for each genres. For

this study the researcher used the top-level page as a seed and then crawled to a depth of three links within the same domain. They chose a depth of three in order to balance the thorough coverage of individual sites with a breadth across many sites. With a depth of three, an average of 6,577 pages for each domain crawled. They collected two main sets of data, first one in May 2005 and then in October 2005 for analysis. From their first sample out of 18 million URLs, 21,200 (around 13.4% of the total sample) instances of spyware were identified. They found that 5.9% of their Web pages were infected by scripted "drive-by download" attacks. Their analysis quantifies the density of spyware, the types of threats, and the most dangerous web zones in which spyware is likely to be encountered. The research also classified different spyware related vulnerabilities on the internet [Moshchuk et. al 2006]. However the results of their research were limited by search depth compared to breadth, as well as the drastic reduction in the numbers of drive by download attack statistics in their consecutive data samples. The presence of web based malware shows how internet users are targeted by infect host of malware, spyware or adware for financial gains.

2.3 Users' Perception towards online privacy

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 [Culnan et al. 2001]. 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 [Culnan, Javelin 2001, 2005]. While most surveys report that a sizable portion of users claim to read such policies or notices regularly [Culnan et. al. 2001], there is evidence to suggest these reports are greatly exaggerated [Jensen et. al. 2005].

Despite legislative efforts, privacy concerns have been shown to be major obstacles to the adoption and success of e-commerce [Adkinson et al 2002]. Numerous surveys indicate that people consider privacy to be important [Belanger, Campbell, Colnan, Earp et. al 1997, 2002, 2000]. The largest U.S. companies do a much better job than their foreign counterparts in putting detailed, meaningful privacy policies on their Web sites compared to their European and Asian counterparts [Cline et. al 2003]. 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 [Jupiter 2002]. A recent study shows some consumers are willing to pay a premium to purchase from more privacy protective websites [Cranor et. al 2007]. Thus 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. Surveys have also found that people are more concerned about their privacy online than offline, even though most cases of identity theft occur offline [Javelin 2005].

2.4 The Platform for Privacy Preferences (P3P)

To overcome some of the problems associated with privacy policies and reduce the burden on users, machine-readable policy specification languages, such as P3P [Cranor et. al. 2004, 2002] and EPAL [Ashley et. al. 2002], have been proposed. The privacy policies can be read by automated agents (such as Privacy bird [Cranor et. al. 2002], Privacy Fox [Arshad et. al. 2004], or the Microsoft IE 6 and 7, or Netscape 7 browsers themselves), and then users are alerted 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. However, it has been found that many privacy languages are available for representing policies, but they tend to use formats convenient to their implementations. There is no single framework or metric to analyze and evaluate the effectiveness of these languages [Ponnurangam

and Cranor 2007]. The impact of this non standardization of privacy languages result in inadequate supports protecting user privacy.

Without question the most popular and widely used of these technologies 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 the privacy policies of the sites [Cranor et. al. 2004, 2002]. P3P specifies a standard XML format for machine-readable privacy policies that can be parsed by a user-agent program. The latest modification of P3P Privacy Finder is an effort to retrieve quick P3P search results from the online privacy policies. These tools have shown some indications of success [Egelman et. al. 2006], though there is only very little data on their effects to show evidence for long-term success or large-scale use for this technology. P3P policies have also been used as data to direct users' web-searches [Cranor et. al. 2003] in a system sharing many methodological similarities to our iWatch. Efforts are being made to standardize the language specification in the P3P 1.1 specification.

2.4.1 Adoption of P3P

To determine P3P adoption among the websites, researchers from AOL and Carnegie Mellon University designed a P3P enabled search engine and gathered some valuable research statistics over the internet on P3P adoption. It is evident from existing current browser technology that users receive information about a particular web site once they submit the HTTP click stream information (IP address, browser version, operating system, etc.) to that web site. For the purpose of not wasting time, users are less likely motivated to visit a different site even after learning about the contents of their privacy policy. Keeping this fact in mind they conducted a study using a modified version of AT&T Privacy Finder. As P3P specifications require that policies remain valid for a period of no less than 24 hours [Cranor et.al 2002], researchers implemented a policy cache along with an improved user interface of privacy finder. P3P –encoded privacy policies associated with the top 20 search results from three search engines were

analyzed based on the queries returned from AOL, Google and Yahoo search engines. They identified that there were two basic problems associated with the P3P user agent design. To run their experiment 19,999 unique search term entered by AOL users were identified. Since most eCommerce sites constantly collect information from shoppers, 25 search terms from Google's Froogle services were also used as input. The sites had embedded P3P policies were evaluated against five APPEL rules. APPEL rules are user preferences stored in a P3P Preference Exchange Language. Researchers analyzed the search results returned by each search engine for each of the search terms, and they found at least one result with a P3P policy for 83% of the search terms. Overall search terms yielded 10% P3P adoption rates. However for eCommerce sites 21% percent P3P adoption rates were found. The top twenty most popular P3P-enabled domains accounted for over 50% of the total number of P3P-enabled hits. A minority of sites from their samples were found to be engaged in direct marketing with or without any way of opting out direct marketing. Even fewer sites were found to share personal information with other companies when the content of the policies were analyzed [Cranor et. al. 2006]. Despites all the efforts many questions remained unanswered. Their results were limited by the fact that many hits came from different pages within a single domain, when all the policies were not unique. In some cases, multiple domain names use the same policies, often because they are owned by the same company or brand name. Their research also did not show statistics about the adoption rate of Compact P3P policies compared to full policies. The full policies were analyzed based on five APPEL rule sets, but the paper did not specify which five APPEL rules were considered, this is essential considering the fact that APPEL rule sets are specific to user's privacy preferences.

2.5 Privacy and legislation

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 [Kunar et.al 2003, Schwartz et. al 1996].

Regardless of approach, the goal of these privacy laws has been 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) , the US Children's Online Privacy Protection Act of 1998 (COPPA) , the US Gramm-Leach-Bliley Financial Services Modernization Act of 1999 (GLBA) and the European Union Directive on the protection of personal data (95/46/EC) [EU Privacy Directives].

Given that studies have shown that users fail to read sites' privacy policies [Meinert et. al 2006], 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 [Anton et. al 2004]. Several privacy groups such as the Electronic Privacy Information Center ("EPIC") and the Center for Democracy and Technology ("CDT") play vital roles in pointing out corporate information privacy and data breaches. Based on the instances and limitation of Fair Credit Reporting Act and the Privacy Act of 1974 (FCRA) "A Model Regime of Privacy Protection v. 2.0" was proposed which is still work in progress [Somoji et. al 2007].

2.6 Privacy Seal Programs

Some companies attempt to self-regulate by adopting privacy 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) which allow licensees who abide by posted privacy policies and/or allow compliance monitoring to display the organization's seal of approval on their web site. Statistics shows that more than 2,000 companies are paying up to $13,000 per year to display these logos on their Web sites. These statistics are valid mostly for ecommerce sites. However it had also been that observed most tech companies don't find it necessary to put seals on their websites. Almost half of the top 50 most-visited websites, display some type of privacy seal. Roughly 7% of Fortune 500 companies display a seal and half of them are in the technology sector. According to 2003 statistics, TRUSTe has nearly a 2-to-1 edge over BBBOnLine in terms of seal adoption program for the top 50 Web sites, and a 3-to-1 edge among Safe Harbor members. However, Better Business Bureau's had higher name recognition (93%) among internet users because of its 91-year history compared to 6-year-old TRUSTe which claims a 69% recognition rate [Cline et. al. 2003]. These self regulation programs have been found to significantly increase consumer trust [Miyazaki et al. 2002], some questions remain over whether what they imply matches user expectations. In other words, there is some indication that users are being misled by some of these efforts [Meinert et. al 2006].

2.7 Privacy Tools

In recent years various privacy protecting tools have been designed to protect user's privacy. In an effort to manage to improve the quality of policy rules, and enable those rules to be implemented through technology to ensure consistency, reliability, and compliance researchers in IBM have designed a tool named SPARCLE. This tool is capable of natural language parsing of privacy policies.

Different anti-spyware tools such as Ad-ware Standard Edition, Aluria Spyware Elliminator, HijackThis, SpyBot-Search & Destroy SpyStopper, SpySweeper,

SpywareBlaster, SpywareGuard, X-Cleaner Free etc are an effort in this direction. Different anti-spam tools such as Anti-Spam Sheriff, Blubottle, MailWasher Norton Internet Security, Outlook-Spam- filter, SpamButcher, SpamNet, Spamtrap, SpamWasher, SpamX etc are also developed to recent years as a measure to protect privacy vulnerabilities via spam [http://www.privacy.gov.au/internet/tools/#spam].

Often pop-ups cause privacy threats to users. These days different advertising filters also used as a pulg-ins to stop pop up windows which often use cookies to collect information from users. These filters includes Ad Block (Powerfull ad blocking plugin for Firefox) AdSubtract, Anti-PopUp for IE, Junkbuster, Mozilla (Advanced Opensource web browser with built in ad and pop-up blocking) Meaya Popup Killer, Muffin, Norton Internet Security, PopUpCop, PopUp Stopper, Privoxy Proxomitron, STOPzilla, Zero Popup Killer etc [http://www.privacy.gov.au/internet/tools/#4].

In recent years, Bugnosis, a web bug remover has gained much popularity as an anti web bug tool [http://www.privacy.gov.au/internet/tools/#5].

As cookies are often used to track user activities different cookie removers has gained substantial popularity. Some of the cookie management and deletion programs such as CM DiskCleaner, Cookie Cop2, Cookie Pal, HistoryKill, IEClean, Norton Internet Security, Tracks Eraser, Web Washer etc are extensively used as a cookie remover to protect user privacy [http://www.privacy.gov.au/internet/tools/#4].

Besides these privacy protecting tools, firewalls also play important role in stopping unauthorized access of the user's machine when they are online. Some popular firewalls are Zone Alarm, Firestarter, NetDefender, Norton Personal Firewall 2006, Black Ice Defender etc [http://www.privacy.gov.au/internet/tools/#3].

Researchers have also designed anti phishing tools and different analysis tools in an effort to reduce the user's privacy concern.

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 effectiveness of such systems, an early shortcoming of many. Classic papers[Tyger, J.D et al.1999, Weirich et al.2001] and studies include 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 [Dhamija et. al. 2006], and how tools and warning tend to go unheeded, regardless of the information presented [Wu et. al. 2006]. Content-based approach to detect phishing web sites and false positive has also shows some promising results in this area [Hong et al. and Cranor et al. www. 2007]. While excellent results, researcher believed that more work still needs to be done in these area as there are far more studies were conducted of why things fail rather than how to succeed.

3 iWatch Background

The iWatch project was initiated at the Georgia Institute of technology in 2004. The project was carried out by Personal Policy Research group of Georgia Institute of Technology to investigate how people think about privacy online [Taberner et. Al 2004]. The aim of this project was to design more efficient, reliable privacy aware tools for the end-users to support their privacy. The major challenge for this project was to provide security and privacy controls for end-users in a dynamic and pervasive computing environment for the future.

3.1 Research Goals at Georgia Institute of Technology

The project was designed as an application of goal-oriented engineering methods to create a rational model of browsing which can analyze security vulnerabilities and propose some principle set of browsing features which can identify a set of privacy vulnerabilities. These vulnerabilities include failure to implement Fair Information Practices (FIP's) [Taberner et. al 2005]. These include accessibility of terms which are on the consolidated list of consumer transactions, Choice and consents, Warranties and Guarantees, Recourse and Redress [Clarke et. al 2006]. The project was designed in the context of Privacy Aware Browsing (PAB) which is an augmented domain that includes browsing mitigating privacy vulnerabilities. However, researchers at the Georgia Institute of Technology derived a certain trade off for finding and countering vulnerabilities by identifying goals [Potts et. al]. Though PAB supported a number of vulnerabilities, Integrity/Security and Redress were not considered and included in the system as it requires protocol and platform modifications to adapt to the system [Taberner et al. 2004]

The initial fundamental research approach for this project was two-fold, developing technologies and approaches which will allow end-users to provide end-users the

capability to better manage and monitor their own personal information when they go online and to develop tools which transform raw data into information about concepts such as centrality or prestige which are widely used in social networking.

3.2 User Study at Georgia Tech

Based on the above research goals, researchers at the Georgia institute of Technology conducted a study to understand users' perceptions about privacy [Jensen and Potts et. al. 2005] Like other surveys this study shows that users are concerned about protecting their privacy, but at the same time it also points out users' lack of perception about privacy aware technologies and their lack of understanding towards them. Their survey report shows that about 90.3% of participants were concerned about the uses of web cookies in the web pages as an user tracking activity, but when investigated thoroughly, only 14% of the participants were actually found knowledgeable about web cookies. In an effort to gather more reliable statistics and implement Privacy Aware Browsing, scholars at Georgia Tech developed an automated web crawler named "iWatch", which follows links to visit web sites and the contents of the pages searching for forms, cookies and other objects and stored them in database.

As privacy is a fluid process, socially negotiable, and constantly evolving, static settings seem to be largely inefficient. Privacy is not just a matter of support which satisfies the end-user's decision making, but it is a complex and important process in terms of e-commerce and its growth as well as the preservation of the social balance among users with the ubiquity of technologies. Hence it involves consumers, legislators, e-Merchants, developers and system administrators and researchers to play their parts and share their responsibilities to add real value in the process.

3.3 Trade-offs of Privacy Aware Browsing

Based on our initial data gathered from the 2005 user studies, we realized the limitation

of our research not involving the legislative part in the Privacy Aware Browsing (PAB), thus we decided to continue this project in a different context, fine tuning the existing architecture of the crawler so that we could gather a reliable set of statistics, collecting data from publicly visible pages of websites over a period of time. As the impact of legislations enforcing redress and recourse plays a vital role for machine privacy policies, we decided to simplify the existing architecture of the crawler to gather large volumes of private data. Our aim is to provide a proof of concept as a part of data and statistical analysis which can be treated as a benchmark for future web analysis.

3.4 Re Initiation of iWatch

One of the principal decisions for reinitiating this project is to show what kinds of data and analysis are possible with even a simple straightforward web crawling methodology. Privacy is an inter-disciplinary process, involving a mixed bag of users, researchers have decided to investigate the effect and impact of legislation on data practices for protecting users' privacy across the world. This requires a reliable set of data samples, over a period of time, which can be used to gauge the evolving nature of privacy practices across the internet landscape. The spread of the dataset needs to be large and representative across the globe, and the data samples should be large enough to have a meaningful analysis. Automated analysis of web using crawl methodology is a very trivial but effective technique for this purpose. However, to gather information about web use and to represent accurate privacy results a large crawl is required over a number of domains.

Though large corporations like Google, and Yahoo are perfectly capable of doing this, given limited university research settings and resources, it is difficult to achieve. The web crawler needs to perform a breadth-wise search to gather accurate information; considering these limitations one of our motivations for this project was to derive a mathematical model which can be used as a base case to determine the depth of crawl

needed from an individual domain. As depth of crawl is proportional to the number of pages that need to be sampled from each domain, any improvement in this direction will be useful for future research and gathering data.

4 Hypothesis and Scope of Analysis

Based on the literature review we asked several research questions, including:

RQ1- Is web crawling a useful methodology for data collection to analyze Internet privacy?

We modified the existing architecture of the crawler in Georgia Tech and ran several experiments using our web crawler. Our scope for performing these experiments was vastly limited with the fact that we have collected our data set within the publicly visible domain of Internet. We did not have a chance to collect data from behind the firewalls. Among all these experiments we have considered three of the main experimental results and showed their evolving nature and how trends of privacy practices have evolved over the three consecutive years.

RQ-2- What kind of privacy analysis is possible using web crawling in the Internet landscape?

Once the datasets for 2005 and 2006 were collected, we analyzed our data and have shown how and to what extent websites are using different correlations between harmful tracking of user activities such as 3rd party cookies; 1st party cookies; web-bugs with different privacy indicators, such as Compact P3P; Full P3P different privacy seals, such as TRUSTe; BBB Online; etc. We have also compared usages across sites of these privacy practices.

RQ-3- How to eliminate geographic bias in web-crawling?

This basically answers our questions what approach we took to improve the bias and what techniques we have used to perform a limited sampling of pages in a meaningful manner. Adding geographic proportions in an efficient and meaningful manner is our answer to this question. Using this methodology we have been able to reduce the over

representation of domain from specific countries and improved the overall bias representation to certain extent.

RQ-4-How to limit the number of pages need to crawl within a domain to save resources?

It is not possible to crawl all the web pages within a domain as web crawling is expensive in terms of resources and bandwidth. That is why we took a probabilistic approach to determine how many pages we need to sample with a maximum probability to determine the usage of privacy vulnerabilities within a specific domain. Once we determine this, we can crawl to some other websites. In this way we can achieve much larger breadth compared to depth.

One of the main limitations we have for evaluating these two techniques are the insufficient number of experimental samples. We only tested the solutions for improving the bias for 2007 data samples and probabilistic model we tested theoretically but it needs further simulation.

5 iWatch Architecture

iWatch is a web-crawler, or spider [Heydon et. al. 1999], 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. While performing this process the crawler can also log these pages or perform other operations on pages fetched according to the requirements of the system. The purpose of the iWatch Crawler is to examine web-pages and keep a history of privacy practices carried out by examined web-pages.

5.1 Crawling policies

Web crawling is difficult to perform in general, due to the large volume of content on the internet and the ever changing features of web sites. Due to the large volume of web pages web crawler can only download a fraction of the pages within a given time. Thus it needs to prioritize which page it should download first. Also the high rate of change means that while a website is being crawled pages might get added or deleted and the crawler needs to adjust at the runtime with these changes.

As the bandwidth for conducting crawls is neither infinite nor free it is essential to crawl the web in a not only scalable, but also efficient [Edwards et al.2001]. A crawler must carefully choose at each step which pages to visit next. The behavior of a web crawler is the outcome of a combination of policies: a) A selection policy that states which pages to download b) A re-visit policy that states when to check for changes to the pages c) politeness policy that states how to avoid overloading websites d) a parallelization policy that states how to coordinate distributed web crawlers. These are

the four main tasks which every crawler needs to perform. Based on these definitions we took a multiphase design approach to make our crawler scalable and efficient.

5.2 Initial Design at Georgia Tech

iWatch was first designed back in 2004 at Georgia Tech. The earlier version of iWatch functions as follows:

First iWatch connects to the database. It then processes the list of URLs of URL from the existing table called SITESTACK and starts crawling. As the architecture is purely multi-threaded, each thread crawls either a new URL or sometimes same URL even, thread writes back to the database after getting new links. Once the new URL has been added to the database, each thread of iWatch processes new requests from the databases. The links are stored in a process queue in the database and are processed as a First In First Out (FIFO) order. The early design of iWatch consisted of several crawler threads: a URL handling thread, a URL packet dispatcher thread and URL packet receiver thread. Each thread picks up an element from the URL pending list, generates an HTTP fetch requests, gets the page, parses through this page to extract any URL's in it and finally puts them in the job pending queue of the URL handling thread. The thread gets a URL from the job queue, checks to see if the URL belongs to the URL set corresponding to the crawler entity. Thread Synchronization is one of the most challenging parts of this implementation.

The most popular 50 URLs are stored in SITESTACK table of iWatch Database to start the crawl. Each thread starts crawling from these URL lists of SITESTACK table. iWatch uses BFS (Breadth First Search) for making decision of its crawling i.e. as a part of its Selection policy. Breadth First Search is being performed with the contents of table SITESTACK along with a join with the table called URL_HISTORY which contains the pages crawled from the URL link based First In First Out order. This

implies the root of the URL sites will be written first in the SITESTACK table first and it will be removed first from the table also.

5.3 iWatch Implementation Iterations

At Oregon State University we tried various approaches in order to make our crawler more robust and stable. We tried to implement High Priority Queue abbreviated as (HPQ) with an idea to distribute the traffic of the crawler. We tried to route the normal HTML pages into the SITESTACK and popups, scripts etc into the HPQ[Appendix-2]. We also tried to implement reading of filters from a separate XML file in-order to clean up the code. Each of these implementations added efficiency to a certain extent, but enforces certain restrictions also. We also modified the earlier database schema to certain extent.

5.4 Current iWatch Working Structure

Final Design of iWatch

Finally we implemented a much simpler design. Though we kept our approach fairly simple we added some post-processing features. These include the addition of an IP mapping capability based on the IP addresses of the websites, mapping the data according to geography, designing separate custom applications for finding and downloading full P3P policies. During each implemented version we tested our code with a different number of threads in order to test the stability and robustness of the crawler.

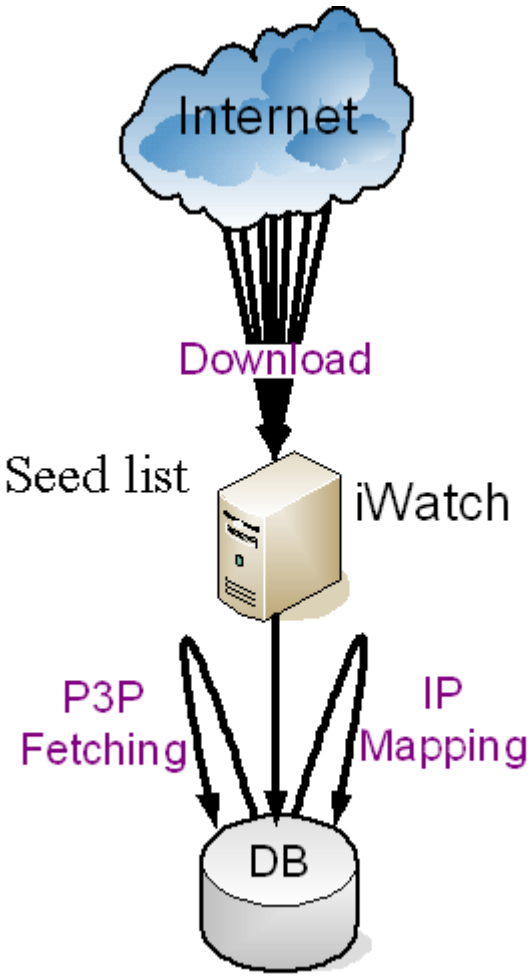


Figure 1: Modified iWatch Design (SOUPS-2007 Slides)

iWatch is a multi threaded web-crawler, or spider [Heydon et al 1999], implemented in Java, and built from the ground up to search for and index data-handling practices. Because of the built in functionality in Java to support network programming the design and the implementation of iWatch was preferred in Java. The implementation in Java also allows to support platform independency for iWatch. iWatch is capable of start crawling independent of operating systems and hardware. Java implementation of iWatch enables us to use and modify our regular expressions directly derived from the

regex class of Java, which are primarily responsible for indexing pattern matching of the different privacy keywords from the HTML pages on the internet.

iWatch search 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) for accurate statistics.

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 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.

iWatch uses MYSQL server 5.0 as a database. The major advantage of using MYSQL as a database is that it supports GUI-based architecture. The MYSQL Server window helped us to monitor load, traffic conditions over the internet, and the status of individual threads. Each iWatch database consists of the following tables: tbl_domains, tbl_seedlist, tbl_sitestack, tbl_urlhistory, tbl_events, tbl_global. The detail of the table structures are presented in the Appendix-A.

One of the major successful modifications in our current version of iWatch is the implementation of GEO IP mapping. As each thread crawls different pages of a URL it registers the IP Addresses for the corresponding pages using specific Java IP Mapping functions. The IP Addresses were logged against specific domains as well as under specific pages in our iWatch database within the URL_HISTORY table. These IP Addresses were later mapped in terms of geographic spread for our datasets.

iWatch uses MaxMind's GeoIP Country database to map the data sets based on the IP addresses of different countries as a post processing strategy. The MaxMind's GeoIP Country database is capable of determining the Internet visitor's country based on the IP addresses. It contains the following fields: Start IP Address, End IP Address, Start IP Number, End IP Number, Country Code ,Country Name. Based on the Start IP Address and End IP Address, crawled domains within a specific country code and country names are mapped.

As a part of post processing of the data, the latest version of iWatch also incorporates a custom Full P3P detection application. The finding and downloading of full P3P policies from the web sites are error prone because filters need to work accurately to eliminate false positives. As specified earlier [Cranor et. al 2003] some servers at times refused to serve full P3P policy downloads from default locations (http://server/w3c/p3p.xml); the custom application revisited crawled domains 3 consecutive times searching for full policies. The repeated queries gave us much more accurate statistics in terms of finding the adoption of full p3p policies.

The last post data processing step of iWatch tries to determine the usage of privacy seals by cross-referencing different seal providers' data lists with crawled domains. iWatch attempts to determine seal usage directly from the pages crawled initially by using its filters. This proved to be an ineffective strategy. As seals 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 the sample sites.

Based on the above architecture, iWatch started gathering data starting from an initial seed list. Because the initial seed-list used has a tremendous effect on the overall crawling pattern, so it is important to choose the seed list 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.

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 of 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.

Filter Name	Filter Description
Cookies	Identifies the use of different types of cookies (session, normal and 3rd party)
Popups/Unsolicited popups	Identifies the use of unsolicited popup windows
Web-bugs	Identifies the use of third part resources potentially used to track users from site to site
Image reorder	Identifies image reordering and hiding, sometimes used to place web-bugs
Banners	Identify the use of different types of banners and adds, potentially used to track users from site to site
Full P3P	Identifies the use of full P3P privacy policies by site
P3P compact policy	Identifies the use of compact P3P privacy policies by site
Crude-parental	Crude parental filter looks for list of curse and pornographic words
Hidden forms	Looks for hidden forms sometimes used to pass along information without the users knowledge
Refresh tags	Identifies refresh tags sometimes used to redirect users or pass hidden information to websites
HTML annoyances	Identifies practice typically associated with predatory sites
Jumping windows	Identifies practice typically associated with predatory sites
IE-exploits	Identifies the use of known Internet Explorer exploits
Javascript annoyances	Identifies different types of known javascript exploits and practices typically associated with predatory sites
Shockwave/flash	Identifies the Macromedia Shockwave or Flash
Quicktime/kiosk mode	Identifies the use of quicktime and quicktime kiosk mode

Table: 1 list of iWatch Filters (Jensen et. al 2005-A)

One of the important functions of iWatch filters is the capability of catching Java script enabled code, and user contributed snippets for unsolicited popups. Though in our current implementation we did not build the functionality for content control our filters are well capable of catching popups generated by third party domains. Recent research on the presence of malware shows how java script-enabled snippets infect web pages. Though our present version of filters are not robust enough to catch third party widgets, user contributed content and web server security for java script snippets in the web pages are capable of catching script-enabled unsolicited popups set by third party domains.

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. The number of pages in each domain in crawl can be controlled by limiting the specific number of pages. This number is hand tuned, and can be increased or decreased.

5.5 The Result of iWatch Crawler

The iWatch Crawler's task is to explore a given set of websites in order to feed a database with data regarding the use of user privacy information managed by the site explored by the crawler.

The relevant information that it should obtain includes the following: is the interaction method between the website and the user Get or Post and different privacy vulnerabilities such as uses of different types cookies, potential harmful webbugs set

from 3rd party domains, Popups set up to identify user tracking activities from 3rd party domains, other java script related privacy vulnerabilities.

6 Important Definitions

Before diving into the experimental methodology, 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, which are chosen for practical considerations.

In this thesis, 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. 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 thesis 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 has been used interchangeably in this thesis.

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 thesis, 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 have tracked 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 sometimes present threat to end-users privacy 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 to some extent.

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 thesis are detected 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 parties.

Some of the 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.

Based on these definitions, we defined our key search indices. These search indices are as represented as follows:

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	Identifies 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/)

Table 2: Abbreviated list of iWatch search indices (Jensen et al. 2007)

Appendix 3 through Appendix 6 has been formatted such that each row in the individual tables for 2006 and 2007 (Full Sample, Filtered Sample) are considered as a single data point. Each data point represents a single country and the word data point has been used consistently through out this thesis.

7 Experimental Methodology

To demonstrate the effectiveness and value of this approach to the study of online privacy, online regulation, and online data collection practices, we performed several experiments. Based on our initial experimental results we modified our experimental procedure continuously to obtain accurate results. We build our first significant analysis based on the experiment conducted in Georgia Tech in May of 2005, and then we conducted our second experiment in August 2006. From both of these data-sets, we analyzed information on web-sites' privacy and data-collection practices. We finally collected our third and most recent sample in May 2007. 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 three samples to examine the changes that have taken place online over the last two year.

7.1 Seed-list

Like other crawler's iWatch starts its operation with an initial seed list, the seed-list has a tremendous effect on the overall crawling pattern. Hence it is important to choose it 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 first two experiments, the crawler was seeded with a combination of the top 50 most popular 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.

Due to the dynamic nature of internet, any two samples are likely to deviate significantly in terms of the sites visited. If the deviation takes place early enough in the crawling 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.

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.

7.2 Modified 2007 Seed-list

Based on the assumption of geographic enforcement and earlier experience, we modified our seed list for the collection of May 2007 Samples. In order to reduce the English language bias and acquired more balanced crawl we chose two domains each from the top 20 countries represented in our 2006 samples. Each of these two domains were chosen based on the prior experience of the researcher as well as based on the recommendation from Google search. As half of the countries in our earlier two samples were predominantly from the most net-populous nations, and at least for the top 20 countries each represented more than 0.50% of the overall global domain-population from our earlier two samples. Once we exit this exclusive group, quirks and bias are less important, given the small relative size of these countries in the net representation based on this diagnosis the remaining seed list was formed based on the hand picked of Asian, European, Latin American and certain African web sites based on the researcher's experiments. Our 2007 data samples shows a much broad representation compared to our earlier two samples as overall our dataset reached a spread of 133 countries.

7.3 Custom P3P Detection

Described earlier, in our architecture section we wrote a custom application to detect the full P3P policies from the crawled domains. As some sites uses multiple redirects through an authentication server before serving the requested page to the server from where original request was made. Without the appropriate passwords and scripting for automated system to authenticate itself, our crawler was unable to verify it accurately. This is indeed what happens and thus we cannot determine automatically which policy reference file is actually applicable. Each domain in crawl is revisited by 3 times trying

to get a full p3p policy by this custom application. These repeated queries made a significant difference in our results, giving us an additional 117 policies out of 1790 full policies (7.65 %) for our 2006 sample, and 211 additional policies out of 2765 full policies (6.53 %) in the 2005 sample when compared with a single visit strategy. Responses were analyzed to check that whether the returned documents were xml document or just a html document, and that redirects were followed correctly. In the current version of the crawler, the P3P policies are not analyzed.

7.4 Privacy Seal Detection

As described earlier because our initial approach to detect the privacy seals directly from the webpages proved ineffective as crawler needs to search for the policy page. We only got 48% detection success for our 2005 samples and 55% detection success for our 2006 samples. 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. We also analyzed our sample against Better Business Bureau Reliability Seal to find out how many of the domains (most eCommerce domains) are consumer reliability concerned.

In this thesis the first two sample results are presented together which will be followed by some of our interesting 2007 dataset results. Given the large size of our all three samples, finding statistical significance is relatively simple even for relatively small changes in behavior.

8 Results

8.1 Combined Analysis of 2005 and 2006

Our combined 2005 and 2006 samples represent a total of 240,340 web pages from a total of 26,213 domains. On average 9.17 pages were analyzed per domain. Table 3 summarizes the basic characteristics of the two samples [Jensen et al. 2007].

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 U.S. websites (Figure 2 shows an overview of our geographic reach). Many of the countries were represented by an 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).

The EEA is composed of the 25 European Union (EU) members, plus Iceland, Liechtenstein, and Norway. All domains belonging to an EEA country were included in our sample because all EEA countries are signatories to the EU privacy directive [EU Privacy Directives], 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 used 27 in our sample (Lichtenstein being absent, see table 3 for a list of all countries included in the study). EEA countries make up 9.66% of our total sample.

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

Table 3: Data Summary for our combined 2005 and 2006 samples

Applying the above filtering rules, we lose 56 domains and 26 countries from Sample 1, and 34 domains and 17 countries from 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 according to Webhosting.info (http://www.webhosting.info/domains/country_stats). This means that though our samples only reached approximately 0.037% of all registered domains, these samples are representative of a large percentage of the net.

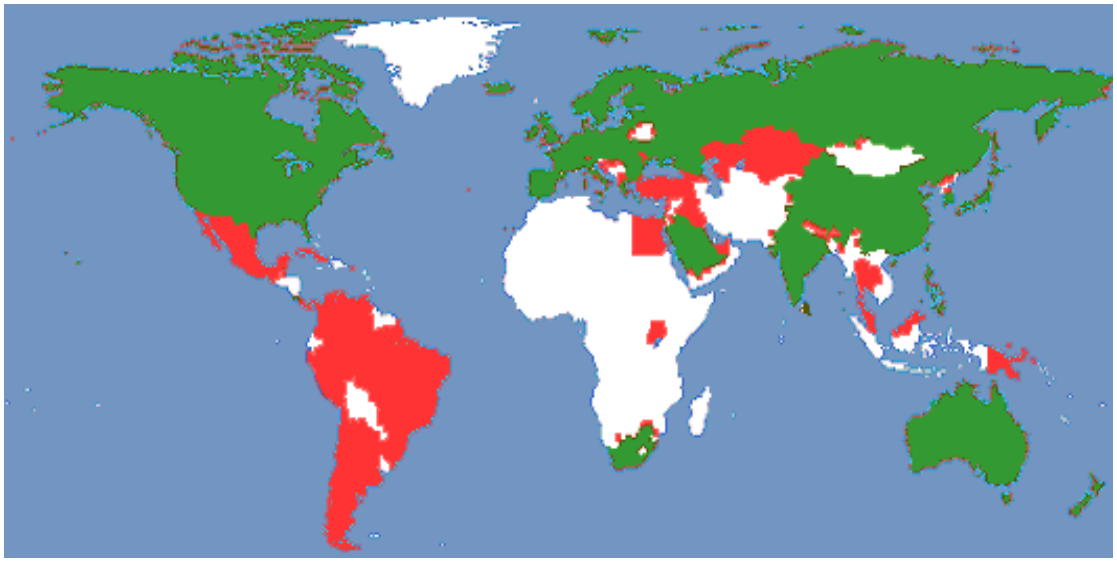


Figure 2: Geographic distribution of combined sample for 2005 and 2006

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

8.2 Global Privacy Practices for Combined 2005 and 2006 Samples

In an effort to identify the trends of global data practices we looked several interesting practices from our 2005 and 2006 Data sets. The Table: 4 represent the percentage of domains adopting practices, and the geographic spread of these practices as percentage of all countries in our sample. Based on a test of proportions a * with green highlight indicates statistically significant increase from one year ago ($P<0.001$)

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: Global data-practices (Jensen et al 2007)

One of the main conclusions from these above figures 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 [Culnan et al. 2001] 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 pop-ups and web-bugs is flat from a year ago, though geographic distribution is up.

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 3). 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 3, popularity does indeed affect the adoption of P3P, though much more markedly today than in 2005.

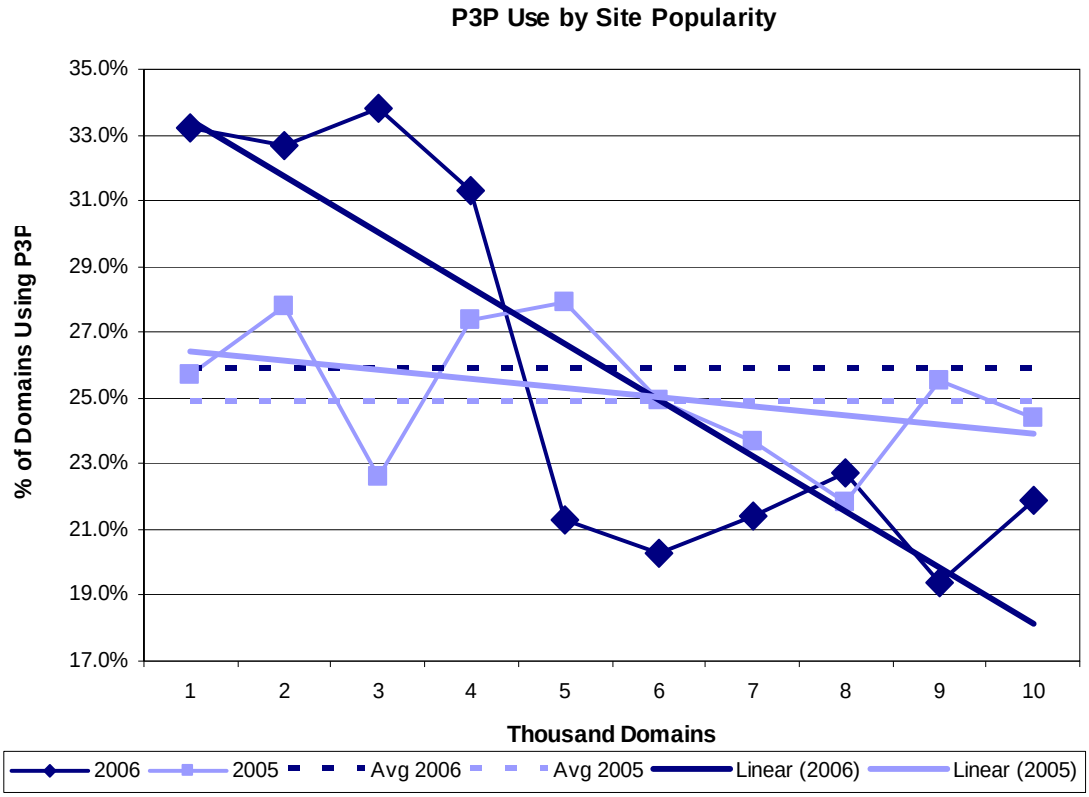


Figure 3: P3P use by site popularity

Figures 4 and 5 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 10 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 3 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.

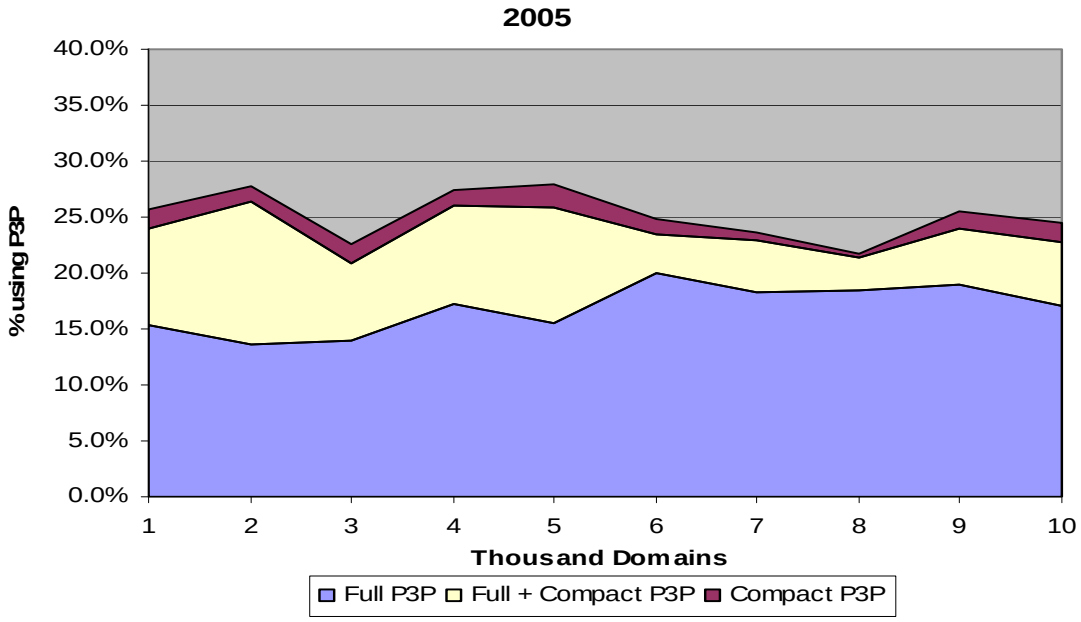


Figure 4: P3P use by site popularity and type, 2005

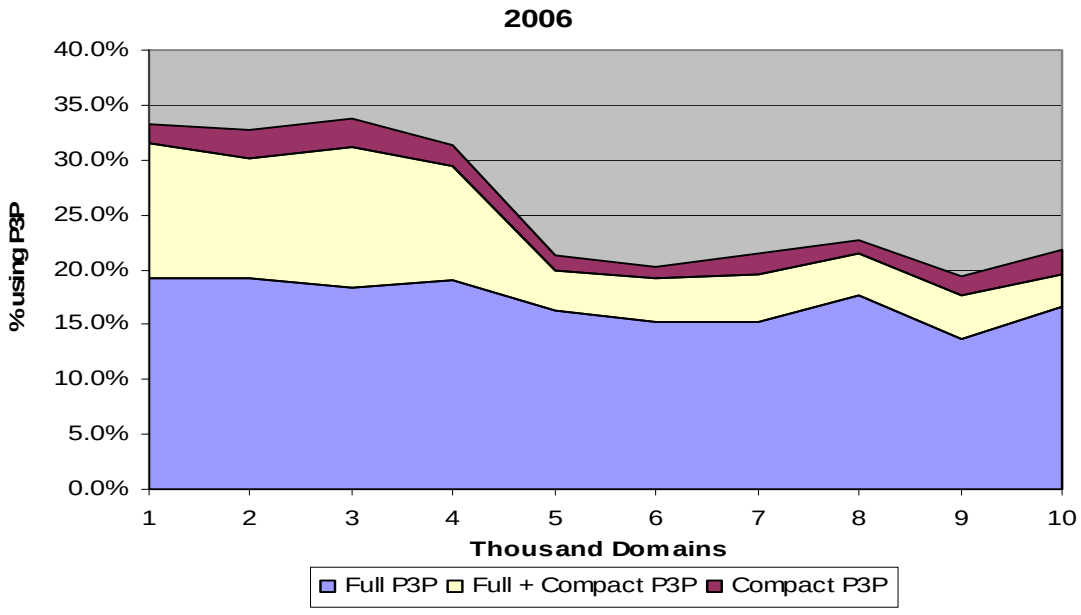


Figure 5: P3P use by site popularity and type, 2006

8.3 Effects of P3P and Privacy Seals on practices

Table -5 shows Effects of P3P and Privacy Seals on practices. Table shows percentage 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 are 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 are 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%

Table 5 Effects of P3P and Privacy Seals on practices

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 web-bugs, 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.

8.4 Effects of BBB Reliability Seals on practices

We also analyzed the adoption rate of Reliability seals for Better Business Bureau (BBB)in our sample. Reputation and other customers' recommendation play major roles in decision making for consumers to start an online business with any website. It has been found that 73% of purchasers and 82% of non-purchasers expressed lack of reliability as a major concern when shopping online in United States (http://www.bbbonline.org/reliability).Reliability seal from BBB addresses the reliability of e-merchants to a certain extent towards consumers. As reliability of online shopping associated with safe data protection, retention policies and towards the safeguard of consumers' privacy, despite of having an U.S centric business approach we have added the adoption of BBB Reliability seals in this thesis.

Table-6 shows correlations with some of the potential harmful user tracking activities along with BBB Reliability seals. Since BBB Reliability seals are mostly complaints to eCommerce sites, we found that strong correlations exists between some of the user tracking activities with BBB Reliability seals in our data set. It is evident from the Table-6, theses co-occurrences are in increase from 2005 to 2006.

	Canada		United States	
	2006	2005	2006	2005
BBB Reliability	0.98%	0.52%	1.06%	0.54%
BBB Reliability+Popups	0.98%	0.26%	0.59%	0.35%
BBB Reliability + Webbugs	0.49%	0.52%	0.17%	0.34%

Table-6 Effect of BBB Reliability seals on Privacy Practices in U.S and Canada

On further investigation with CHI SQUARED TEST ($p<.001$) , we have found that there is a significant statistical increase in uses of webbugs and popups with the reliability seal in our dataset from 2005 to 2006 for United States

Webbugs: $X_1^2 = 106.9968343$ and two sided P value=4.45881E- 25

8.5 Impact of legislation on Data Practices

One of the intended uses of these two data-sets is to examine the effects that legislation and regulation have on data-practices. Table 7 gives an overview of the geographic clustering of data.

	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)

Table 7: Geographic clustering of domains

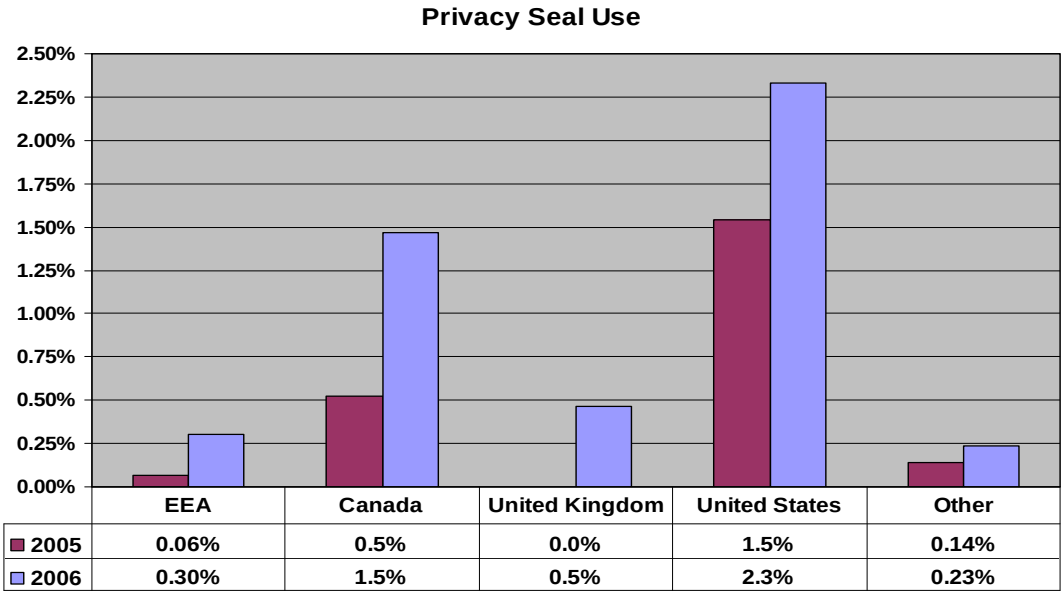
Table shows number of countries and the % of all domains in each group and sample. In the total column we give the actual number of domains. * UK appears both on its own and as part of the EEA sample. Based on a test of proportion, cells marked by *or

with green or tan highlight in 2006 Detected column indicates statistically significant increase or decrease from one year ago ($p<0.01$).

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.
As mentioned earlier, 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 for 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.

8.6 Global trends

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. Figure 6 and Figure 7 represents the similarity of adoption of U.K, Canada and United States in terms of Privacy seal and 3rd Party Cookies.



	EEA	Canada	United Kingdom	United States	Other
2005	0.06%	0.5%	0.0%	1.5%	0.14%
2006	0.30%	1.5%	0.5%	2.3%	0.23%

Figure 6: 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 ($p<0.005$)

Some of the interesting observations are that, as Figure 6 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.

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.

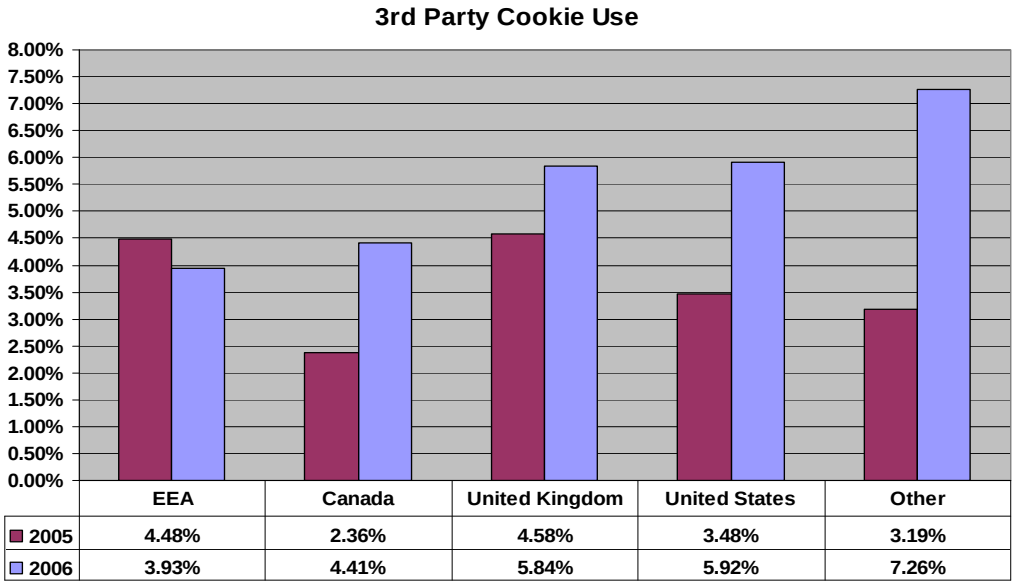
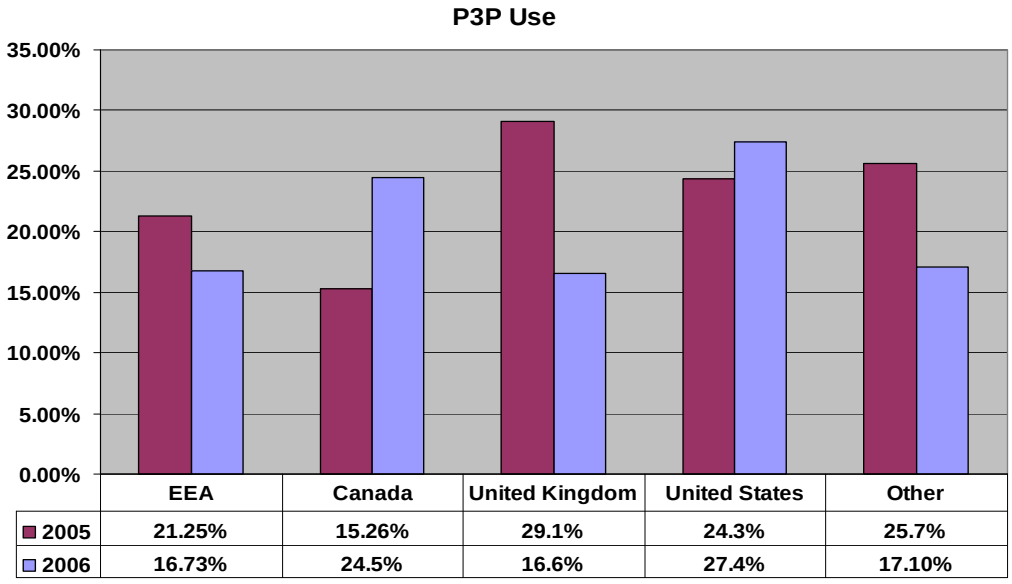


Figure 7: 3rd Party Cookie use by geographic area

Horizontal bars showing global average for the two samples (by color). All changes from 2005 to 2006 are statistically significant ($p<0.005$).Clearly the adoption of 3rd party cookies for U.K and Canada and U.S follows identical pattern.

Another interesting finding is the skew in P3P adoption, with the US and Canada very much leading the way (Figure 8), 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. One of the probable reasons might be, since U.K is a part of EEA which follows “EU Privacy Directives” British websites are less inclined to use P3P as an indicator to display the data collection strategies for the users.



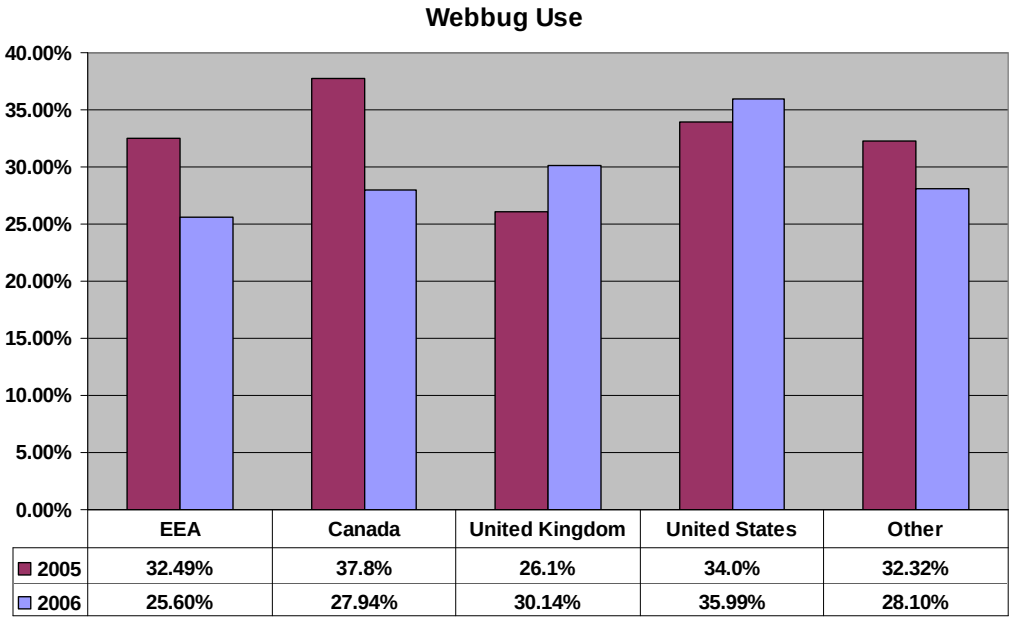
	EEA	Canada	United Kingdom	United States	Other
2005	21.25%	15.26%	29.1%	24.3%	25.7%
2006	16.73%	24.5%	16.6%	27.4%	17.10%

Figure 8: 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 ($p<0.005$)

While other technologies could have been examined in this fashion, we decided to conclude this study by looking at the problematic web-bugs which is a potential threat to end-users privacy.



	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 9: Web-bug use by geographic area

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

In terms of adoption of web-bugs it is evident U.K is following the trends of U.S we found a reverse trend for Canada when we compared our data-set from 2005 to 2006. While the observed trends were in line with our expectations, the differences were not as marked as we had expected, nor were they uniform, thus requires further investigation in near future.

8.7 Presence of Bias in Combined 2005 and 2006 Samples

Table 8 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 8), including most EEA countries (highlighted in light grey in Table 8). Some smaller countries, through quirks, 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 negotiation efforts were receiving widespread international press.

Country	Total		Samples		Bias
	Number of Domains	% of Domains	Number of Domains	% of Domains	(% of expected)
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 8: Bias in the combined samples in 2005 and 2006 (Jensen et al. 2007)

Based on a test of proportions, the * and # symbols in the bias column together with green and tan highlighting, respectively, indicates significant positive or negative bias ($P<0.001$). It is evident from the above table that, given the large size of our two samples, finding statistical significance is relatively simple even for relatively small changes in behavior. We therefore uncharacteristically within the field of computer science, chose to set our threshold for statistical significance at the $p < 0.001$ level.

8.8 Combined Result of analysis for 2005, 2006 and 2007

Our combined 2005 and 2006 data samples were collected from a seed list which was a combination of Comscore Media Matrix's list for that specific year and a hand-picked set of popular Asian and European websites. The seed list was constructed in such a way that it maximizes the potential value or impact on the overall dataset. A fully representative sample requires random sampling, which is not possible with a web crawler as it investigates clusters of websites by following the links between them. In an effort to rectify this over representation, we ensured that the most popular sites, the sites most likely to impact the privacy of users, were at the heart of the crawl. Since our seed list was U.S centered, our sample showed a heavy bias towards U.S. websites. As a result of this, some of the countries were represented with only very small number of domains. Table 8 shows, the bias in our combined 2005 and 2006 sample was statistically significant at the $p < 0.001$ level for approximately half of the countries, this bias is less than what we had expected from heavily Internet-active nations. Once we exclude this exclusive group, quirks and bias were less important, given the small relative size of the countries. For instance, Norway was over-represented with 223 domains, 436.47% of the sample size, which only accounted 0.85% of the overall sample size.

Based on the earlier experience, we modified our seed list for the collection of the May 2007 Samples. As stated briefly in our methodology section in order to reduce the English language bias and acquire a more balanced crawl we chose two domains each from the top 20 countries represented in our 2006 samples. We took the top 20 countries these two domains were chosen based on the prior experience of the researcher from our earlier experiments as well as based on the recommendation from Google keyword search. As half of the countries in our earlier two samples were predominantly from the most net-populous nations, each represented more than 0.50% of the overall global domain-population from our earlier two samples. Once we exit this

exclusive group, quirks and bias are less important, given the relative small size of these countries in the net representation based on this diagnosis the remaining seed list was formed based on the hand picked set of Asian, European, Latin American and certain African web sites based on the researcher's experiments. Our 2007 data samples shows a much broad representation compared to our earlier two samples as overall our dataset covered 133 countries.

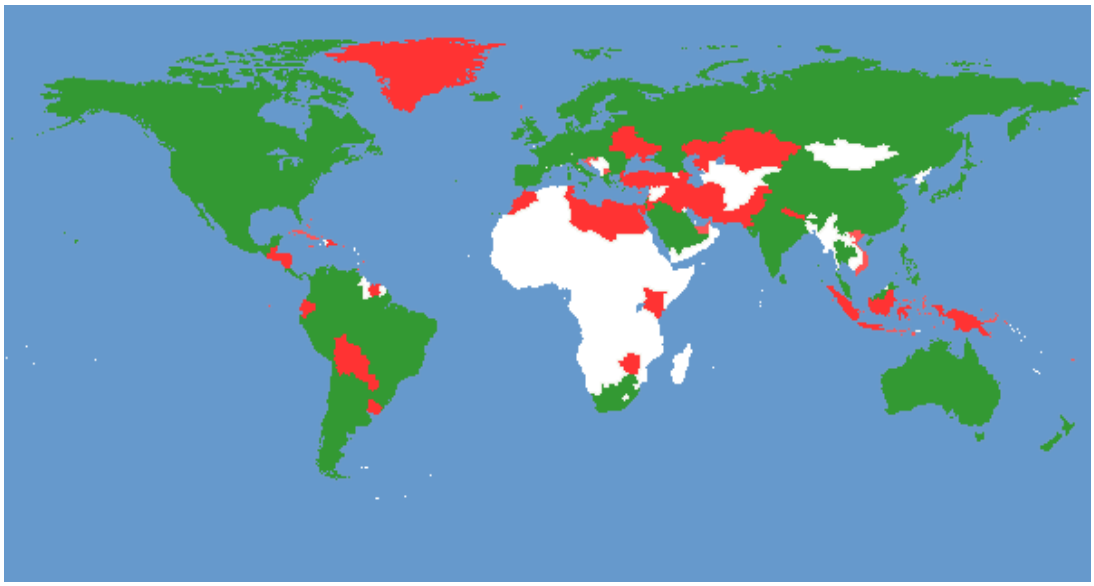


Figure 10: Geographic distribution of combined sample for 2005, 2006 and 2007
Countries marked in green are included in the study. Countries marked in red were reached, but excluded from the study due to small sample size. Map courtesy of world66.com

Across the 3 samples, a total of 618,860 web pages were crawled from a total of 53,528 domains. This represents 11.56 web pages/domain. Given the speed with which websites evolve, and that these samples were taken 24 months apart, we decided to use the non-unique total in our calculations and to treat the three samples as statistically independent. This means that on average we analyzed 11.56 webpages per domain,

which is 2.6 web pages/domain more than our earlier analysis (Jensen el al. 2007), a relatively solid basis for drawing conclusions about any given domain.

Overall, our three samples reached 117 countries or territories: 69 in the first sample, 60 in the second, and 106 in the third sample (Figure 10 shows an overview of our geographic reach). In our combined 2005 and 2006 sample, many of these countries were represented by an extremely small number of domains and pages in our data-sets, which forced us to filter some of the data. We applied a similar filtering rule but much stricter manner. If any country which is not represented by more than 10 domains are excluded from the analysis, unless they were part of the European Economic Area (EEA). EEA countries make up 17.43% of our total sample.

	Sample 1	Sample 2	Sample 3	Total
Collection	May 2005	August 2006	May 2007	
Web-pages	119,237	121,103	377,728	618,068
Domains	15,792	10,421	27,392	53,605
Web-Pages/Domain	7.55	11.62	13.78	11.53
Total Countries	69	60	106	117
Filtered Countries	43	43	59	60
Domains/Country	367.26	242.35	462.27	893.42

Table 9: Data sample summary statistics for 3 samples

Applying the above filtering rules, we lose 148 domains and 58 countries from our total sample. Overall, 57 countries were filtered from the combined data-set, leaving 60 (43 in each from the first two samples, 59 from third sample). Table 9 shows on average, the excluded countries were only represented by 3.96 domains in our sample. As could be expected, our probes primarily reached the most net-active countries in world. Though we only saw a total of 60 countries, those countries account for more than 94% of all active domains according to Webhosting.info

(http://www.webhosting.info/domains/country_stats). This means that though our samples only spread approximately 0.06% of all registered domains, these samples are representative of a large percentage of the net.

Table 10 shows the distribution of domains across countries, as well as the bias of the sample relative to the countries' current (August 2007) Internet footprint. It is evident, that our 3 years of consecutive samples contain bias; however, we have been able to reduce the bias compared to our combined samples of 2005 and 2006. This has been achieved by adding definite geographic proportions in our seed list. Based on experience from our earlier experiments we have modified our earlier seed list. Our seed list comprises of 75 domains for this purpose. We have selected 2 domains each from the top 20 countries from Table 4: which basically represents 95% of the overall Internet population. We have also added 3 domains from Africa, 3 domains each from Latin America and the Middle East, and one from New Zealand so that our crawl can spread geographically. All these domains are selected based on given keyword search in Google. As the United States represents the highest percentage in the overall net population, 23 domains were selected from four prominent categories of web such as: eCommerce, education, social networking, and financial websites. We have also added 2 Chinese websites in the seed list for greater geographic spread in Asia.

Country	Number of Domains	% of Domains	Number of Domains	% of Domains	(% Expected)
United States	55,599,209	66.58%	38158	71.29%	*107.07%
EEA	14554920	17.43%	6162	11.51%	#66.05%
Germany	4,849,319	5.81%	1046	1.95%	#33.65%
United Kingdom	3,257,454	3.90%	2031	3.79%	97.27%
Canada	2,855,737	3.42%	1533	2.86%	#83.75%
China	2,558,685	3.06%	3532	6.60%	*215.36%
France	1,934,042	2.32%	670	1.25%	#54.05%
HongKong	1,692,536	2.03%	417	0.78%	#38.44%

Australia	1,644,771	1.97%	435	0.81%	#41.26%
Japan	1,081,247	1.29%	875	1.63%	*126.25%
Spain	1,016,797	1.22%	630	1.18%	96.66%
Korea	874,285	1.05%	345	0.64%	#61.56%
Italy	811,091	0.97%	254	0.47%	#48.86%
Netherlands	661,621	0.79%	361	0.67%	85.13%
New Zealand	495,427	0.59%	34	0.06%	#10.71%
India	408,859	0.49%	199	0.37%	#75.94%
Russia	353,202	0.42%	269	0.50%	118.82%
Denmark	297,265	0.36%	122	0.23%	#64.03%
Sweden	253,173	0.30%	176	0.33%	108.46%
Brazil	245,124	0.29%	64	0.12%	#40.73%
Norway	216,599	0.26%	349	0.65%	*251.38%
Austria	207,989	0.25%	80	0.15%	#60.01%
Switzerland	200,016	0.24%	189	0.35%	*147.42%
Poland	189,725	0.23%	40	0.07%	#32.89%
Belgium	133,320	0.16%	78	0.15%	91.28%
Mexico	126,102	0.15%	27	0.05%	#33.40%
Thailand	122,487	0.15%	15	0.03%	#19.11%
Ireland	118,114	0.14%	60	0.11%	79.25%
Finland	113,087	0.14%	80	0.15%	110.37%
Czech Republic	107,214	0.13%	43	0.08%	62.57%
Argentina	102,096	0.12%	106	0.20%	*161.98%
Bulgaria	98,213	0.12%	14	0.03%	#22.24%
Malaysia	83,070	0.10%	16	0.03%	#30.05%
Israel	75,837	0.09%	530	0.99%	*1090.33%
Portugal	62,916	0.08%	12	0.02%	#29.76%
Singapore	62,333	0.07%	128	0.24%	*320.37%
South Africa	57,452	0.07%	53	0.10%	143.92%
Taiwan	48,514	0.06%	99	0.18%	*318.37%
Hungary	40,858	0.05%	18	0.03%	68.73%
Romania	39,575	0.05%	22	0.04%	86.73%
Panama	36,495	0.04%	11	0.02%	47.02%
Venezuela	35,865	0.04%	14	0.03%	60.90%
Saudi Arabia	35,132	0.04%	41	0.08%	*182.07%

Colombia	33,488	0.04%	19	0.04%	88.52%
Greece	32,913	0.04%	25	0.05%	118.50%
Luxembourg	31,814	0.04%	22	0.04%	107.89%
Peru	29,458	0.04%	15	0.03%	79.44%
Philippines	28,738	0.03%	27	0.05%	146.58%
Estonia	23,710	0.03%	3	0.01%	19.74%
Slovenia	23,635	0.03%	15	0.03%	99.01%
CostaRica	20,055	0.02%	69	0.13%	*536.77%
Chile	15,727	0.02%	34	0.06%	*337.28%
Latvia	12,631	0.02%	6	0.01%	74.11%
Slovakia	12,007	0.01%	4	0.01%	51.97%
Lithuania	11,898	0.01%	9	0.02%	118.01%
Malta	11,230	0.01%	1	0.00%	13.89%
Sri Lanka	6,988	0.01%	45	0.08%	*1004.67%
Gibraltar	6,819	0.01%	2	0.00%	45.76%
Georgia	3,565	0.00%	52	0.10%	*2275.66%
Iceland	3,526	0.00%	4	0.01%	176.99%
Sample Total	**83,511,055**	**94.18%**	**53528**	**100.00%**	
Global Total	**88,667,541**				

Table 10: Bias in the combined samples in 2005, 2006 and 2007

Table 10 represents the improved percentage of positive and negative bias in the combined three samples. Comparing with our combined 2005 and 2006 samples, it has been found that adding geographic proportion has improved the percentage of bias for top 20 countries. As an example, we can say that the positive bias for United States has been decreased by 16.87. This result is statistically significant with X_1^2=836.6014022 and P value= 5.9495E-184 based on CHI Squared Test (P<0.001); Similarly negative bias for EEA has been increased by an amount of 14.53% which is statistically significant based on our CHI Squared test with a (P<0.001) as X_1^2= 22678578956 and P Value= 0.00. Also the bias for individual EEA countries also improved to a certain extent. When we compared our combined 2005, 2006 and 2007 samples with our combined result sets of 2005 and 2006, we found that our strategy was successful for 14

out of top 20 countries in our combined samples. Adding geographic proportions in the seed list has decreased the amount of bias except for China, Japan, India, New Zealand, Switzerland and Austria. The addition of 4 Chinese domain, and 2 each from both Japan and India have added a cumulative positive bias of 265.67% with X_5^2= 1879.247954 and P Value= 0.00 CHI Squared Test (P<.001) which we failed to realize when we constructed the seed list. As the net population of Switzerland and Austria represents only 0.24% and 0.25% respectively compared to overall population, taking two domains from Austria and Switzerland and representation of other European domains in the seed list probably introduced bias in the sample. The number of registered domains has drastically increased from 2006 to 2007 in New Zealand; choosing only one domain in the seed list resulted in a negative bias in our combined result.

8.9 Individual Spread and Geographic Bias 2006 Sample vs. 2007 Sample

Based on the above results, we decided to further investigate our 2006 and 2007 samples. As stated earlier, since all EEA countries are signatories of EU Privacy directives we decided to treat them same. Appendix 3 and Appendix 4 represent the overall spread of our 2006 samples and 2007 Samples. In our 2006 Samples, domains from the United States were over-represented (124.97%) and as a consequence other countries were represented in a small number. After adding geographic proportions we found that the US was underrepresented (90.19%) and representation of other countries increases. We have found, significant improvement in representation of U.S based samples to 34.78% based on CHI Squared Test (P<.001) having X_1^2= 266.1724 and P Value= 7.74705E-60. This was our main aim: to modify the seed list to increase geographic representation. Similarly, underrepresentation of EEA countries has improved to a extent to 40.17%, based on CHI Squared test (P<0.001) having X_1^2= 200.3126025 and P Value= 1.78491E-45.This includes the increase in the representation for most of the EEA countries such as Germany, U.K, France, Spain, Italy, Denmark, Sweden. Besides EEA, an increase in the percentage of representation

also takes place for countries like Canada and Australia. However, enforcing geographic proportion in the seed- list also increases the amount of overrepresentation for smaller countries, as in the case of countries like Spain, and Sweden with an increase of positive bias of 127.79% [X_1^2= 25.38598624 and P Value= 4.69315E-07] and 138.09% [X_1^2= 11.870431 and P Value= 0.000570322] respectively, based on CHI Squared test with a (P<0.001). However as these countries are represented by much smaller number of domains compared to United States, so over representation of these countries should have lesser impact in the samples compared to over representation of United States.

8.10 Results after applying Filtering Rule to 2006 and 2007 Samples

Though our initial investigation shows some amount of success, we decided to further investigate this methodology using a filtering rule of 10 for 2006 and 2007 samples unless they are part of EEA. By filtering rule we mean that if any country is not represented by more than 10 domains we have excluded from our analysis. Applying a filtering rule of 10 for individual samples in 2006 and 2007 reduces the size of our data points considerably, which is why we have been able to investigate our samples thoroughly. We found 28 countries remain in our 2006 samples and 57 countries in 2007 samples after applying the filtering rule. When we calculated the bias in both the samples, we found that 16 data points in 2006 samples and 22 data points in 2007 samples contain bias (Appendix-5 and Appendix-6). This amounts to an overall 3.51% [X_{209}^2= 1037.489606 and P Value= 3.2589E-151] improvement in bias for all the data points from the 2006 sample to the 2007 sample based on CHI Squared test with a (P<0.001), confirming our earlier conclusion that applying geographic proportions was a successful strategy.

8.11 Probabilistic Model

One of the principal limitations of automated web crawling methodology is the lack of ability to precisely control the distribution in the sample. We have already seen that despite these limitations, this process is increasingly gaining popularity among researchers. In our literature review we have shown that two papers of this year, one from researchers from Google and other one from University of Washington used web crawling techniques, where they have crawled a tremendous amount of pages with a particular domain. However, within limited university setting it is very expensive to crawl to such a huge depth. That is why we felt that if this automated web crawl methodology is to gain further popularity, there needs to be a model that can predict the depth of pages needs to be crawled to find the privacy and security vulnerabilities.

8.12 Our Methodology to determine the Inputs for the model

Our initial of design of this probabilistic model came when we carefully looked in detail about our data. As for example we found that our crawler has crawled 24 pages from a domain under United Kingdom named "www.atlarge.com" and found cookies in 2 pages.

So this gives us the initial probability of randomly finding a cookie on this site as being 2/24= 0.083

So probability of not finding a cookie will be (1-0.083) = 0.917

So if we sample 5 pages probability of not finding a cookie will be (0.917) ^ 5= 0.6484

Hence the probability of finding a cookie = (1-0.6468) = .3532

So if we sample 5 pages within that domain there is only 35% chance of getting a cookie from that 5 pages. However this calculation is valid when the number of pages is infinite as for example for very large web sites like CNN.com.

Keeping this calculation is mind we clustered our datasets identifying the two most potential privacy vulnerabilities; Cookies and Web bugs, from domains which in our sample contain more than 1, 2, 10, 20 pages in our sample. We then plotted the presence of cookies, web-bugs and P3P based on the number of pages within a domain crawled by the crawler. We compared the results for each of these clusters to see the nature of the distribution and spread., as an example we are analyzing the nature and distribution of cookies and web-bugs from our 2007 data samples which contains more than 20 pages crawled within each domains.

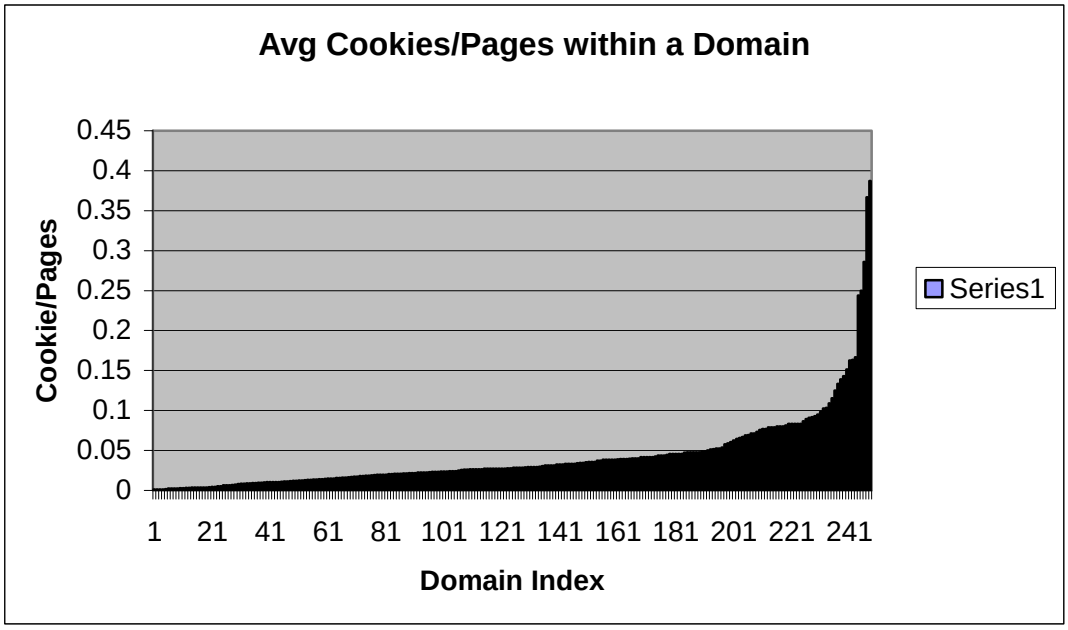


Figure 11: Average Use-Cookies/Pages for different domains

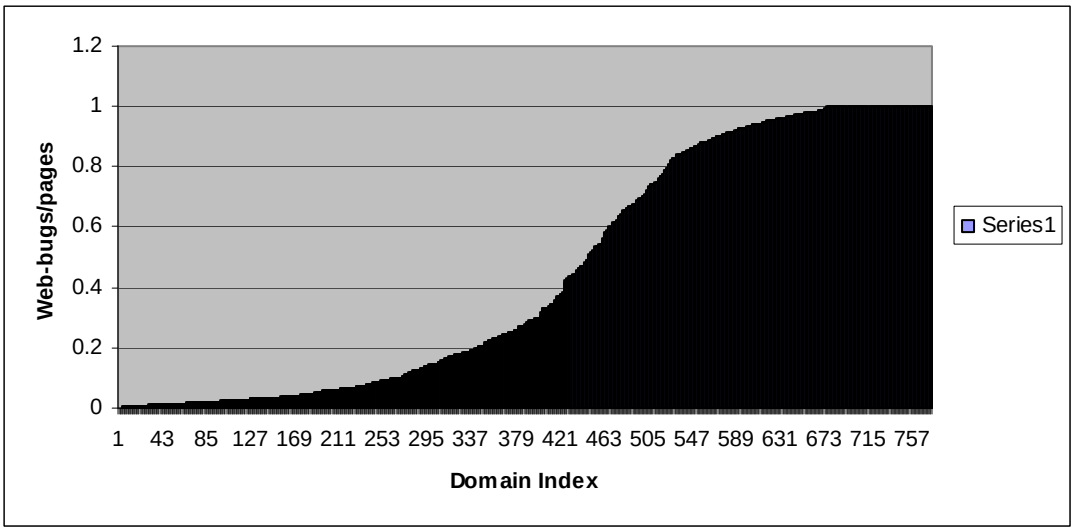


Figure 12: Average Use-web bugs/Pages for different domains

Similarly Figure-7 represents the presence of usage of web-bugs/pages for different domains. Out of 1486 domains, we found 775 domains contain web-bugs set by third parties. From the plots for average number of cookies and average number of web-bugs in different domains we can think, the pages needs to be sampled to find a specific percentage of occurrence of a cookie or webbugs with a domain.

8.13 Our Proposed Probabilistic Model

Our aim for web crawls was to gather information on the internet; we believed that any mathematical model which can predict the number of pages that need to be sampled to detect privacy vulnerabilities will save resources in terms of space, memory as well as band width and speed for this automated analysis, and these saved resources can be effectively utilized to crawl more domains.

Suppose we want to detect how many pages we need to sample/crawl in order to find probability of occurrence of 90% of cookies within a specific domain, assuming crawler has initially crawled Y=20 number of pages and number of cookies in those Y pages were initially X=2.

Let us consider, the probability of finding a cookie in x number of pages is P(C).

$$P(C) = X / Y = 2 / 20 = 0.1$$

Hence, probability of not finding a cookie will be

$$P(\bar{C}) = (1 - X/Y)$$

Clearly, the probability of not finding a cookie after two trials is going to be

$$P(\bar{C}1) = (1 - X/Y) * (1 - X/Y)$$

Let, t to be the number of trials required to find a cookie with 90% probability

Case- 1: Assuming the number of pages in the domain is infinite (appropriate for very large web sites such as cnn.com, yahoo.com etc).

After t trials/samples probability of not finding of a cookie can represented by

$$P(\bar{C}c) = (1 - x/y) * (1 - x/y) * (1 - x/y) \ldots .\text{t times}$$

Let

$$(1 - x/y) = u$$

$$P(\bar{C}c) = u\text{^}t$$

Hence,

$$\therefore P(\bar{C}c) = u\text{^}t = \varepsilon$$

$$\Rightarrow t = \log \varepsilon / \log u$$

Where

$$\varepsilon = (1 - 0.90) = 0.1$$

So the number of pages needs to be sampled for this page will be log(0.1)/log(0.9)=21.85=22

Clearly 5 pages need to be sampled in that domain to find a probability of 40% chance of a cookie.

Case-2: When the number of pages is finite and small, say for mom and pop websites. This case can best be represented using Hypergeometric Distribution [http://mathworld.wolfram.com/HypergeometricDistribution.html]to estimation of how many trials we need for this case

Let us consider the total number of pages for a small domain to be N. The web crawler has crawled y pages and detected x number of cookies.

Clearly the initial sampling probability of occurrence of a cookie in a page is x/y.

This x/y needs to be accurate in order to estimate the number of pages correctly. Let, n be the number of pages need to be sampled to find one cookie of 90% of probability According to Hypergeometric distribution

The probability of i successful selection can be estimated by

P(x=i) = [# ways for i successes]*{[#ways for N-i failures]/[total number of ways to select]}

$$= \frac{\binom{n}{i}\binom{m}{N-i}}{\binom{m+n}{N}}$$

$$= \frac{m!\,n!\,N!\,(m+n-N)!}{i!\,(n-i)!\,(m+i-N)!\,(N-i)!\,(m+n)!}.$$

From our initial experiment if the experimental probability is

P=x/y= Sampling Probability

then for N pages total in the domain we should have at least (Sampling probability* N) cookies which is N*x/y= N'

Let us assume we need to scan n pages to find at least 1 cookie with 40% probability.

Let X = # of pages has cookies out of the t pages.

P(X=x) = (N' Cx)(N- N' Cn-x)/(NCn)

Now

P(X=0) = (N' C0) (N- N' Cn)/(NCn)

Now P(X >= 1) = 1-P(X=0)

= 1-{ (N' C0) (N- N' Cn)/(NCn)} ----------i)

Now in order to determine the number of pages needs to be sampled it should satisfy the inequality of

P(X >= 1) >0.90

1- (N' C0) (N- N' Cn)/(NCn) > 0.90

Since N' and N are known the only unknown is n which can be derived by repeatedly calculating different values of n which satisfies the above inequality i).

Now say let us consider a small websites which contain N= 250 total pages.

Assuming initial experimental probability of occurrence of cookie in a page is

x/y=2/20

This x/y must be accurate in order to estimate the number of pages we need to scan to get one cookie with 90% probability.

In this case we need to use Hypergeometric distribution to approximate this estimation.

According to Hypergeometric distribution

Here P=1/10

For N pages we should have at least (Sampling probability* N) cookies.

Let us assume we need to scan t pages to find at least 1 cookie of 40% probability.

Let X = # of pages has cookies out of the t pages.

P(X=x) = (25Cx)(250-25Cn-x)/(250Cn)

Now

P(X=0) = (25C0) (250-25Cn)/(250Cn)

Now P(X >= 1) = 1-P(X=0)

= 1- (25C0) (225Cn)/(250Cn)

Now we have to calculate repeatedly to solve for different values for n which holds the inequality

1- (25C0) (225Cn)/(250Cn) > 0.90

n = 21

21 pages needs to be sampled for this case.

9 Conclusions

The goals of this thesis 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.

We have found a 36% overlap between our 2005 and 2006 seed-list. Even though our samples were only separated by a year, this likely would have lead to a significant divergence of the 2005 and 2006 two samples, and possibly false inferences about changing practices, if the sample site were too small. With a large enough sample size, all things should even out. We took a different approach to construct our 2007 seed-list, however when compared we found 17 URL's are common in 2005, 2006, 2007 seed-list.

This brings us to the question of whether a sample size of 0.02% of all domains in our combined analysis of 2005 and 2006 and 0.06% for our 2005, 2006 and 2007 samples are 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 4 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 [Flake et al. 2003]. 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. However our methodology to reduce the over representation of bias is successful to considerable extent.

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 researchers' 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 web-bugs. 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 web-bugs.

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. A

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. Earlier research has shown that [Moores et al. 2003], 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.

We are also greatly pleased to improve the over representation of our bias for the 2007 sample and we expect using our derived mathematical model researchers will be usable to use this technique to gather reliable data set on the internet in the future.

Bibliography

[Cranor et al. 2007] L, Cranor. J. Tsai, S. Egelman, and A. Acquisti. The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study. Paper presented at the Workshop on the Economics of Information Security, June 7-8, 2007, Pittsburgh, PA

[Westin et al. 2001]A. Westin. Opinion Surveys: What Consumers Have To Say About Information Privacy,. Testimony before U.S. House of Representatives, Committee on Energy and Commerce, Subcommittee on Commerce, Trade, and Consumer Protection, May 8, 2001.

[Bennett et al. 1992] C. Bennett Regulating Privacy: data protection and public policy in Europe, Cornell University Press 1992, P-263

[Anton et al. 2007] A. Anton. Testimony before the House Committee on Ways and Means Subcommittee on Social Security on Protecting the Privacy of the Social Security Number from Identity Theft 21 June 2007 http://waysandmeans.house.gov/media/pdf/110/Anton.pdf

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

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

[Smith J. et al. 1993] S. Milberg, J Smith, S Burke. This article consists of 23 page(s). Information Privacy Concerns, Procedural Fairness, and Impersonal Trust: An Empirical Investigation Organization *Science*, Vol. 10, No. 1 (Jan. - Feb., 1999), pp. 104-115

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

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

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

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

[Jensen et al. 2005] C. Jensen., C. Potts. "Privacy Policies as Decision-Making Tools: A Usability Evaluation of Online Privacy Notices" *Proceedings of CHI'04* Vienna, Austria, April 2004

[Adkinson et al. 2001] F. Adkinson, A. Eisenach, , M., Lenard. *Privacy Online: A Report on the Information Practices and Policies of Commercial Web Sites*. Progress and Freedom Foundation, Washington DC. March 2002.

[Jensen et al. 2007] C. Jensen, C, Sarkar, C, Jensen, C, Potts. Tracking Website Data-Collection and Privacy Practices with the iWatch Web Crawler Soups, July 18-20, 2007 PA

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

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

[Arshad et al.2004] "Privacy Fox - A JavaScript-based P3P Agent for Mozilla Firefox." *Privacy Policy, Law, and Technology*. 17-801

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

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

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

[Cranor et al 2003]L, Cranor., S, Bayers, D, Kormann, “Automated Analysis of P3P-Enabled Web sites” Proceedings of the 5th International Conference on Electronic Commerce, ICEC2003

[EU Privacy Directives] 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.

[Jensen et al. 2005] C.Jensen, Potts, C., and C Jensen “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.

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

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

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

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

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

[Taberner et al. 2005] L,Taberner, http://www.dsic.upv.es/~einsfran/pfc/PFC-FI-Leandro.pdf

[Clarke et al. 2006] R Clarke: A major impediment to B2C success is...the concept 'B2C'. ICEC 2006: 441-450

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

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

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

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

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

[Jensen et al. 2005-A] C, Jensen."Designing for Privacy in Interactive Systems", A Dissertation Presented to The Academic Faculty College of Computing Georgia Institute of Technology December 2005 http://etd.gatech.edu/theses/available/etd-11272005-044459.

Appendices

Appendix-1

Different Level of iWatch Implementation

The main purpose of the design of iWatch Crawler is to examine web-pages of different websites and keep history of privacy practices carried out by examined web-pages to provide recommendation and solutions of different privacy sensitive issues to the end users.

The different level of implementation which I did in this term for this iWatch project is as follows:

Making the iWatch Crawler live

It was not at all easy to give a live running shape of the earlier existing crawler, out of at least more than ten different versions. As there were no specific documentations were present of earlier versions it was kind of hard for me to understand the earlier code and made modifications based on that. But after initial struggle finally we were able to compile and run several versions of earlier crawler.

Reading and Writing of Filters from Separate Files

Our earlier version of crawler had several hard coded filters defined in the program iWatchCrawler.java. Based on these filters each new thread crawls. These hard coded makes the program very cumbersome and it was difficult to maintain. As maintenance has become a major issue we decided to separate these predefined filters from iWatchCrawler.java into a separate files and read from that so that any modification

requires in these filter. For more detail kindly refer \\spectre\jensenca\Projects\iWatch\documentation\Filter Modifications.doc

Two different levels of implementation

A. Defining the filters in a text file and read the filter from that file for crawling.

B. Defining the filters in a XML file using castor project by using certain marshaller and unmarshaller programs. As marshaller program will generate an XML file test1.xml containing all these filters and unmarshaller program defined in the iWatchCrawler.java will read these filter from test1.xml and perform the crawling.

As xml files are more efficient to maintain we have decided to go with XML based implementation for future work.

Implantation of High Level Priority Queue

Our web crawler iWatch had initially only one level of queue which is tbl_sitestack. The main function of this queue is to add new links to the queue as each time crawler finds new links from the web. Clearly the importance of the queue site stack is immense. As along with the new links crawler also loads web bugs, image URL into the queue maintaining traffic has become an issue. So we decided to bifurcate the entire traffic implementing High Priority Queue (HPQ) so that scripts, pop-ups, images, frames are loaded in the high priority queue and only normal web links are loaded in the SiteStack. For more detail kindly refer to \\spectre\jensenca\Projects\iWatch\documentation\HPQ Documentation.doc

These are the specific details in a step by step manner for implementation of HPQ:

A. Created insertIntoHpq () (source iWatchCrawler.java) method for inserting diiferent links to SiteStack table.

B. Modified the findLinks () methods (source iWatchCrawler.java) for the implementation of traffic control as discussed earlier.

C. Modified the meta refreshes and frames (source iWatchCrawler.java) for traffic control.

D. Created function RecordToSelectSitestack() to select all the records from the table SiteStack and from table global which are not present in the SiteStack.

E. Created function RecordToSelectHpq() to select all the records from the table Hpq and from table global which are not present in the SiteStack.

F. Created function SelecttblSitestack() to select all the records from the table SiteStack and from table domain which are not present in the SiteStack.

G. Created function SelecttblHpq()to select all the records from the table SiteStack and from table domain which are not present in the SiteStack.

The implement this High Priority Queue is on the verge of completion .However it is not fully implemented so far we have to make some decision of the distribution of the traffic control based on Http Specification, which will be implemented in the next phase.

4. Version Implementation Detail

Old_Crawler1

This is the latest version of iWatch which I designed and implemented contains the page limiting capacity of a particular domain. The maximum number of pages can be modified by changing the parameter Max_Number_Domains in the initialization.

newiWatch Version

This version of iWatch was designed for the newly designed database new_iwatch2. The database was designed with 5 main tables. In this version tbl_sitestack had been replaced partly with the page_queue and partly with the links. We removed the tbl_global in this version of database design. Url_events and Url_history are the exact aliases of tbl_events and tbl_urlhistory. We used thread.exit() to solve the heap space problem of Java. The crawler worked substantially what we had expected to perform. The thread management for this version was much more optimized.

Crawler db

This was the version, we ran to collect iWatch4 data. This version was the last version which worked perfectly in Oregon State and which had a close resemblance with the earlier versions of Georgia tech. The updataion of tbl_global worked correctly, maxid was fixed, the count in domains worked fine. Only problem we had was out of heap space problems for eclipse or sometimes crawler starved as we run 300-500 threads and its fails find new domains and links after running 3-4 days.

iWatch cralwer version 10

This was the version of crawler we used to gather the iWatch6 data collection. This version of crawler performed proper updates of domain counts as well as the max_id implementation accurately. We optimized the join query in the beginning of the crawl when the crawler was about to start crawling and collecting the data from the tbl_sitestack gathering seed list) .We used regular expressions directly in the instead of reading the regular expression from xml file to eliminate the java io exceptions.

iWatch crawler version 9

In this version of crawler we added the Pagetest code to verify the whether the crawler was collecting the frames and every links within the frames. The pagetest was implemented using http unit.

iWatch crawler version 9B

This version of the crawler, we added frames as well as the links within the pages, we tested this version with having specific links under frames as well as under normal links.

iWatch crawler version 8

This version of crawler we performed some testing, running some test cases using JUnit and JMeter. The idea was to inspect the performance benchmark and code coverage for this version.

iWatch crawler version 7

In this version we modified the implementations standard of Thread management and insertion of records adding specific methods in each case. We implemented HPQ in this version. The idea was here to grab normal html links, java scripts, mailto etc all should be added to tbl_sitestack however links with frames will be populated in the HPQ. The result set will start its query from the HPQ at first and if it's exhausted it will go the tbl_sitestack.

iWatch Crawler Version 6

In this version of the crawler we implemented the filter modification. We read the filters which are basically regular expressions from a XML file using the Marshaller and Unmarshaller interfaces. The real nice thing about this implementation was, text parsing was done automatically. Initially using the Marshaller interface we read all the regular expressions within the crawler and regular expressions were parsed into xml files and after that we read the XML file using unmarshaller import. We used project library castor to implement this part.

N.B: In some of the later version we used the regular expression instead of reading from a XML file because some reading from a file causing java I.O exceptions which was expected as 300 thread are trying to open the same file and read the regular expressions at the same time.

iWatch Crawler Version 5

In this version we implemented the filter modification reading the regular expressions from a text file. We used normal reading from a text file to gather user tracking technologies for data collection. This version was much more simpler than the version 6. However if run 300-500 threads at a time this version also going to have java.i.o exception.

iWatch Crawler Version 1-4

The iwatch Crawler version 1-4 we have used to test to identify the most accurate version we should start using which was live in Georgia tech. Each versions was having their own pros and cons.

Appendix-2

List of Personal Profile identified by the browser cookie

-js_annoyances.

-js_events.

-html_annoyances.

-content_cookies.

-refresh_tags.

-unsolicited_popups.

-all_popups.

-img_reorder.

-banners_by_size.

-banners_by_link.

-tiny_textforms.

-jumping_windows.

-frameset_borders.

-Demoronizer.

-shockwave_flash.

-quicktime_kioskmode.

-Fun.

-crude_parental.

-ie_exploits.

-site_specifics.

Appendix –3

(2006 Full Sample)

Country	Total		Sample		Bias
	Number of Domains	% of Domains	Number of Domains	% of Domains	(% of expected)
United States	46036912	67.61%	8798	84.49%	*124.97%
EEA	12543466	18.42%	710	6.82%	#37.01%
Germany	4039278	5.93%	154	1.48%	#24.93%
United Kingdom	2947932	4.33%	428	4.11%	94.94%
Canada	2495501	3.66%	204	1.96%	#53.46%
China	2099671	3.08%	11	0.11%	#3.43%
France	1733082	2.55%	137	1.32%	#51.69%
Australia	1393853	2.05%	54	0.52%	#25.33%
Spain	884969	1.30%	8	0.08%	#5.91%
Japan	871196	1.28%	68	0.65%	#51.04%
Korea	837088	1.23%	152	1.46%	118.74%
Hong Kong	763480	1.12%	11	0.11%	#9.42%
Italy	721992	1.06%	26	0.25%	#23.55%
Netherlands	547838	0.80%	90	0.86%	107.43%
India	342735	0.50%	22	0.21%	#41.98%
Denmark	263789	0.39%	19	0.18%	#47.10%
Russia	240386	0.35%	22	0.21%	59.85%
Sweden	209208	0.31%	18	0.17%	56.26%
Switzerland	186619	0.27%	26	0.25%	91.11%
Norway	172123	0.25%	10	0.10%	37.99%
Austria	163612	0.24%	15	0.14%	59.95%
Poland	141423	0.21%	6	0.06%	#27.74%
Finland	123288	0.18%	16	0.15%	84.87%
Belgium	122048	0.18%	18	0.17%	96.44%
Czech Republic	91051	0.13%	9	0.09%	64.64%
Israel	81883	0.12%	16	0.15%	127.78%
Bulgaria	81290	0.12%	2	0.02%	16.09%

Ireland	73363	0.11%	17	0.16%	151.53%
Portugal	56850	0.08%	1	0.01%	11.50%
New Zealand	53517	0.08%	5	0.05%	61.10%
South Africa	48384	0.07%	6	0.06%	81.09%
Taiwan	48254	0.07%	17	0.16%	*230.38%
Romania	35479	0.05%	5	0.05%	92.16%
Hungary	31249	0.05%	3	0.03%	62.78%
Greece	27661	0.04%	4	0.04%	94.56%
Philippines	25859	0.04%	5	0.05%	126.44%
Luxembourg	23819	0.03%	1	0.01%	27.45%
Gibraltar	19162	0.03%	2	0.02%	68.25%
Costa Rica	19152	0.03%	3	0.03%	102.43%
Estonia	14640	0.02%	1	0.01%	44.67%
Lithuania	9988	0.01%	1	0.01%	65.47%
Latvia	8332	0.01%	1	0.01%	78.48%
Sri Lanka	5821	0.01%	1	0.01%	112.34%
Sample Total	**68093777**	**96.27%**	**10413**	**100.00%**	
Global Total	**70733538**				

Appendix -4

(2007 Full Sample)

Country	Total		Sample		Bias
	Number of Domains	% of Domains	Number of Domains	% of Domains	(% of expected)
United States	46036912	67.61%	8798	84.49%	*124.97%
EEA	12543466	18.42%	710	6.82%	#37.01%
Germany	4039278	5.93%	154	1.48%	#24.93%
United Kingdom	2947932	4.33%	428	4.11%	94.94%
Canada	2495501	3.66%	204	1.96%	#53.46%
China	2099671	3.08%	11	0.11%	#3.43%
France	1733082	2.55%	137	1.32%	#51.69%
Australia	1393853	2.05%	54	0.52%	#25.33%
Spain	884969	1.30%	8	0.08%	#5.91%
Japan	871196	1.28%	68	0.65%	#51.04%
Korea	837088	1.23%	152	1.46%	118.74%
Hong Kong	763480	1.12%	11	0.11%	#9.42%
Italy	721992	1.06%	26	0.25%	#23.55%
Netherlands	547838	0.80%	90	0.86%	107.43%
India	342735	0.50%	22	0.21%	#41.98%
Denmark	263789	0.39%	19	0.18%	#47.10%
Russia	240386	0.35%	22	0.21%	59.85%
Sweden	209208	0.31%	18	0.17%	56.26%
Switzerland	186619	0.27%	26	0.25%	91.11%
Norway	172123	0.25%	10	0.10%	37.99%
Austria	163612	0.24%	15	0.14%	59.95%
Poland	141423	0.21%	6	0.06%	#27.74%
Finland	123288	0.18%	16	0.15%	84.87%
Belgium	122048	0.18%	18	0.17%	96.44%
Czech Republic	91051	0.13%	9	0.09%	64.64%
Israel	81883	0.12%	16	0.15%	127.78%
Bulgaria	81290	0.12%	2	0.02%	16.09%
Ireland	73363	0.11%	17	0.16%	151.53%
Portugal	56850	0.08%	1	0.01%	11.50%
New Zealand	53517	0.08%	5	0.05%	61.10%

South Africa	48384	0.07%	6	0.06%	81.09%
Taiwan	48254	0.07%	17	0.16%	*230.38%
Romania	35479	0.05%	5	0.05%	92.16%
Hungary	31249	0.05%	3	0.03%	62.78%
Greece	27661	0.04%	4	0.04%	94.56%
Philippines	25859	0.04%	5	0.05%	126.44%
Luxembourg	23819	0.03%	1	0.01%	27.45%
Gibraltar	19162	0.03%	2	0.02%	68.25%
Costa Rica	19152	0.03%	3	0.03%	102.43%
Estonia	14640	0.02%	1	0.01%	44.67%
Lithuania	9988	0.01%	1	0.01%	65.47%
Latvia	8332	0.01%	1	0.01%	78.48%
Sri Lanka	5821	0.01%	1	0.01%	112.34%
Sample Total	**68093777**	**96.27%**	**10413**	**100.00%**	
Global Total	**70733538**				

Appendix-5

(2006 Sample with a filtering rule of 10 except EEA)

Country	Total		Sample		Bias
	Number of Domains	% of Domains	Number of Domains	% of Domains	(% of expected)
United States	46036912	67.61%	8798	84.49%	*124.97%
EEA	12543466	18.42%	710	6.82%	#37.01%
Germany	4039278	5.93%	154	1.48%	#24.93%
United Kingdom	2947932	4.33%	428	4.11%	94.94%
Canada	2495501	3.66%	204	1.96%	#53.46%
China	2099671	3.08%	11	0.11%	#3.43%
France	1733082	2.55%	137	1.32%	#51.69%
Australia	1393853	2.05%	54	0.52%	#25.33%
Spain	884969	1.30%	8	0.08%	#5.91%
Japan	871196	1.28%	68	0.65%	#51.04%
Korea	837088	1.23%	152	1.46%	118.74%
Hong Kong	763480	1.12%	11	0.11%	#9.42%
Italy	721992	1.06%	26	0.25%	#23.55%
Netherlands	547838	0.80%	90	0.86%	107.43%
India	342735	0.50%	22	0.21%	#41.98%
Denmark	263789	0.39%	19	0.18%	#47.10%
Russia	240386	0.35%	22	0.21%	59.85%
Sweden	209208	0.31%	18	0.17%	56.26%
Switzerland	186619	0.27%	26	0.25%	91.11%
Norway	172123	0.25%	10	0.10%	37.99%
Austria	163612	0.24%	15	0.14%	59.95%
Poland	141423	0.21%	6	0.06%	#27.74%
Finland	123288	0.18%	16	0.15%	84.87%
Belgium	122048	0.18%	18	0.17%	96.44%
Czech Republic	91051	0.13%	9	0.09%	64.64%
Israel	81883	0.12%	16	0.15%	127.78%
Bulgaria	81290	0.12%	2	0.02%	16.09%
Ireland	73363	0.11%	17	0.16%	151.53%
Portugal	56850	0.08%	1	0.01%	11.50%

New Zealand	53517	0.08%	5	0.05%	61.10%
South Africa	48384	0.07%	6	0.06%	81.09%
Taiwan	48254	0.07%	17	0.16%	*230.38%
Romania	35479	0.05%	5	0.05%	92.16%
Hungary	31249	0.05%	3	0.03%	62.78%
Greece	27661	0.04%	4	0.04%	94.56%
Philippines	25859	0.04%	5	0.05%	126.44%
Luxembourg	23819	0.03%	1	0.01%	27.45%
Gibraltar	19162	0.03%	2	0.02%	68.25%
Costa Rica	19152	0.03%	3	0.03%	102.43%
Estonia	14640	0.02%	1	0.01%	44.67%
Lithuania	9988	0.01%	1	0.01%	65.47%
Latvia	8332	0.01%	1	0.01%	78.48%
Sri Lanka	5821	0.01%	1	0.01%	112.34%
Sample Total	**68093777**	**96.27%**	**10413**	**100.00%**	
Global Total	**70733538**				

Appendix-6

(2007 Sample with a filtering rule of 10)

Country	Total		Sample		Bias
	Number of Domains	% of Domains	Number of Domains	% of Domains	(% of expected)
United States	46036912	67.61%	8798	84.49%	*124.97%
EEA	12543466	18.42%	710	6.82%	#37.01%
Germany	4039278	5.93%	154	1.48%	#24.93%
United Kingdom	2947932	4.33%	428	4.11%	94.94%
Canada	2495501	3.66%	204	1.96%	#53.46%
China	2099671	3.08%	11	0.11%	#3.43%
France	1733082	2.55%	137	1.32%	#51.69%
Australia	1393853	2.05%	54	0.52%	#25.33%
Spain	884969	1.30%	8	0.08%	#5.91%
Japan	871196	1.28%	68	0.65%	#51.04%
Korea	837088	1.23%	152	1.46%	118.74%
Hong Kong	763480	1.12%	11	0.11%	#9.42%
Italy	721992	1.06%	26	0.25%	#23.55%
Netherlands	547838	0.80%	90	0.86%	107.43%
India	342735	0.50%	22	0.21%	#41.98%
Denmark	263789	0.39%	19	0.18%	#47.10%
Russia	240386	0.35%	22	0.21%	59.85%
Sweden	209208	0.31%	18	0.17%	56.26%
Switzerland	186619	0.27%	26	0.25%	91.11%
Norway	172123	0.25%	10	0.10%	37.99%
Austria	163612	0.24%	15	0.14%	59.95%
Poland	141423	0.21%	6	0.06%	#27.74%
Finland	123288	0.18%	16	0.15%	84.87%
Belgium	122048	0.18%	18	0.17%	96.44%
Czech Republic	91051	0.13%	9	0.09%	64.64%
Israel	81883	0.12%	16	0.15%	127.78%
Bulgaria	81290	0.12%	2	0.02%	16.09%
Ireland	73363	0.11%	17	0.16%	151.53%
Portugal	56850	0.08%	1	0.01%	11.50%
New Zealand	53517	0.08%	5	0.05%	61.10%

South Africa	48384	0.07%	6	0.06%	81.09%
Taiwan	48254	0.07%	17	0.16%	*230.38%
Romania	35479	0.05%	5	0.05%	92.16%
Hungary	31249	0.05%	3	0.03%	62.78%
Greece	27661	0.04%	4	0.04%	94.56%
Philippines	25859	0.04%	5	0.05%	126.44%
Luxembourg	23819	0.03%	1	0.01%	27.45%
Gibraltar	19162	0.03%	2	0.02%	68.25%
Costa Rica	19152	0.03%	3	0.03%	102.43%
Estonia	14640	0.02%	1	0.01%	44.67%
Lithuania	9988	0.01%	1	0.01%	65.47%
Latvia	8332	0.01%	1	0.01%	78.48%
Sri Lanka	5821	0.01%	1	0.01%	112.34%
Sample Total	**68093777**	**96.27%**	**10413**	**100.00%**	
Global Total	**70733538**				

Appendix-7

2007 Seed- list

URL name			
http://www.myspace.com	T		
http://www.google.com	O		
http://www.germany-tourism.de	P		
http://www.bildt.t-online.de			
http://www.bbc.co.uk	2		
http://www.guardian.co.uk	0		
http://www.cbc.ca			
http://www.radio-canada.ca	C		
http://www.sohu.com	O		
http://www.sina.com.cn	U		
http://www.lemonde.fr	N		
http://fr.yahoo.com	T		
http://www.theaustralian.com.au	R		
http://www.ratestogo.com.au	I		
http://www.canalmeteo.com	E		
http://www.marca.es	S		
http://www.wle-japan.com			
http://www.asahi.com	T		
http://www.arakor.co.kr/english/english_AboutArakor.asp	A		
http://www.tour2korea.com	K		
http://www.hongkongpost.com	I		
1878668.socialnet.org.hk	N		
http://www.italyrentals.com	G		
http://www.ferroviedellostato.it			
http://www.loquo.com	2		
http://www.clubpca.com			
http://www.gujarati-online.com	A		
http://timesofindia.indiatimes.com/cms.dll/default	T		
http://www.casinoverdiener.com			
http://www.cmcbiopharmaceuticals.com	A		
http://www.hro.org			
http://www.mapryal.org	T		
http://www.erlang.org	I		
http://www.eniro.com/en	M		
http://www.lepointdufle.net	E		
http://www.swissworld.org			
http://visitnorway.com	FROM		
http://www.telenor.com			
http://www.eurozine.com	2006		
http://www.moondial.com/	Samples		
http://www.nedbank.co.za	South Africa		
http://www.kwv.co.za			
http://www.economatica.com	Portugal		

http://www.mundolatino.org/prensa/			
http://lanic.utexas.edu	Latin America		
http://www.georgetown.edu/LatAmerPolitical/home.html			
www.ddynamic.net	Israel		
http://www.aerotel.com/company.asp?id=14			Middle East
http://www.bnatksa.com/	Saudi Arabia		
http://www.intuto.com/	New Zealand		
http://www.nytimes.com			
http://www.gatech.edu	H		
Http://www.wastington.edu	a		
http://www.theregister.com	n		
http://www.cnn.com	d		
http://www.espn.com			
http://www.classmates.com	P		
http://www.ivillage.com	i		
http://www.cnet.com	c		
http://www.vg.no	k		
http://www.wachovia.com	e		
http://www.citibank.com	d		
http://www.fifa.com			
http://www.olympics.com	S		
http://www.monster.com	i		
http://www.orbitz.com	t		
http://www.ivillage.com	e		
http://www.cnet.com	s		
http://www.citysearch.com			
http://www.untd.com	from		
http://www.hotelmarketing.com			
http://www.adobe.com	U.S		
http://www.whitepages.com			
http://www.online.sh.cn			
http://www.china.org.cn			

Appendix-8

2006 Seed- list

URL Name	
http://www.doubleclick.com	C
http://www.nytimes.com	O
http://www.gatech.edu	M
http://www.msn.com	S
http://www.theregister.com	C
http://www.cnn.com	O
http://www.marca.es	R
http://www.aftenposten.no	E
http://www.myway.com	
http://www.bbc.co.uk	M
http://www.tomshardware.com	E
http://www.aol.com	D
http://www.espn.com	I
http://www.classmates.com	A
http://www.ivillage.com	
http://www.ea.com	M
http://www.cnet.com	A
http://www.symantec.com	T
http://www.microsoft.com	R
http://www.real.com	I
http://www.ebay.com	X
http://www.expedia.com	
http://www.shopping.com	L
http://www.weather.com	I
http://www.gator.com	S
http://www.casino.com	T
http://www.guardian.co.uk	
http://www.bildt.t-online.de	T
http://www.theaustralian.com.au	O
http://www.washington.edu	P
http://www.elcorteingles.es	
http://www.lonelyplanet.com	5
http://www.news.com	0

http://www.abc.es	
http://www.vg.no	W
http://www.loquo.com	E
http://www.lemonde.fr	B
http://www.pravda.ru	S
http://www.cbc.ca	I
http://www.theonion.com	T
http://www.whitehouse.org	E
http://www.export.gov/safeharbor/	S
http://www.wachovia.com	
http://www.citibank.com	
http://www.amdzone.com	
http://www.fifa.com	
http://www.olympics.com	
http://www.dell.com	
http://www.slashdot.com	
http://www.blogger.com	

Appendix-9

Source Code for Custom P3P Detection Application

```
//  Custom P3P Detection Apps
/*
This program detects the presence of compact, full and full and compact policies in our crawled domains. Te detection is done based on HHTP response from the stream and finally detected results are separated in different databases.
*/
import java.io.BufferedInputStream;
import java.io.BufferedReader;
import java.io.IOException;
import java.io.InputStreamReader;
import java.net.HttpURLConnection;
import java.net.MalformedURLException;
import java.net.URL;
import java.sql.DriverManager;
import java.sql.PreparedStatement;
import java.sql.ResultSet;
import java.sql.SQLException;
import java.sql.Statement;
```

```
import java.util.Map;

import com.mysql.jdbc.Connection;

public class P3P extends Thread {

private static HttpURLConnection link;

private static Connection con;

private static URL currentURL;

private static String urlName;

private static Statement stmt;

private static int iThread;

/**

* Constructor that initializes P3P thread.

**/

public P3P(int ThreadNum) throws ClassNotFoundException, SQLException {

iThread = ThreadNum;

}

// connection = con;

public static void main(String args[]) throws ClassNotFoundException {

try {

Class.forName("com.mysql.jdbc.Driver");

con = (com.mysql.jdbc.Connection)

DriverManager.getConnection("jdbc:mysql://localhost/p3poldsub1?user=iwatch&
```

```
password=kec3061");
} catch (SQLException e) {
}
try {
Statement stmt;
ResultSet rs;
String sqlQuery = "SELECT domain FROM bad_domain";
System.out.println("sqlQuery");
stmt = con.createStatement();
rs = stmt.executeQuery(sqlQuery);
rs.first();
// if number of records == 0
if (rs == null) {
return;
} else
 {
while (rs.next()) {
try {
urlName = null;
urlName = rs.getString("domain");
if (urlName.indexOf("http://") != 0) {
```

```
urlName = "http://" + urlName;
}
currentURL = new URL(urlName);
// fetch content from server
link = (HttpURLConnection) currentURL.openConnection();
System.out.print(link.getConnectTimeout());
link.setConnectTimeout(5000);
link.setReadTimeout(5000);
boolean compactExists = false, fullExists = false;
String responseMessage = link.getResponseMessage();
if (responseMessage.equals("OK")) {
Map header = link.getHeaderFields();
/* Checking for Compact policy in the header
 CP="CAO DSP COR CUR ADM DEV TAI PSA PSD IVAi IVDi CONi TELo
OTPi OUR DELi SAMi OTRi
UNRi PUBi IND PHY ONL UNI PUR FIN COM NAV INT DEM CNT STA
POL HEA PRE GOV" Compact policy always exist in the header as a key value
pair where P3P is the key and value is
* CP */
// if (header.containsKey("P3P"))
if (header.containsKey("P3P")||header.containsValue("CP")) {
```

```
compactExists = true;
//          System.out.println(header);
/* List val = (List) header.get("P3P");
String policy = (String) val.get(0);*/
}
}
/*
* If the response message of the url is not O.K(200)  for its going to give a my error
* This will also useful for forbidden message etc
*/
else
{
System.out.println("my error due to bad url links" +urlName);
String query =("insert into P3Poldsubsub1.bad_response(domain) values('" + urlName + "')");
PreparedStatement pstmt=con.prepareStatement(query);
pstmt.executeUpdate(query);
pstmt.close();
}
/*
```

```
* Checking for full policy
* The format of the full policy as specified by w3c: w3c specified
http://host.domain/w3c/p3p.xml
*/
String p3purl = urlName + "/w3c/p3p.xml";
currentURL = new URL(p3purl);
link = (HttpURLConnection) currentURL.openConnection();
if (link.getResponseMessage().equals("OK")) {
fullExists = true;
}
System.out.print("Thread: " + iThread
+ " Crawling url: " + urlName);
if (fullExists && compactExists) {
System.out.println(" - both");
String query =("insert into P3Poldsubsub1.comandfull_policy(domain) values('" +
urlName + "')");
PreparedStatement pstmt=con.prepareStatement(query);
pstmt.executeUpdate(query);
pstmt.close();
}
else if (!fullExists && compactExists)
```

```
 {
System.out.println(" - compact");
String query=("insert into P3Poldsubsub1.compact_policy(domain) values('" +
urlName + "')");
PreparedStatement pstmt=con.prepareStatement(query);
pstmt.executeUpdate(query);
pstmt.close();
}
 else if (fullExists && !compactExists) {
System.out.println(" - full");
String query = ("insert into P3Poldsubsub1.full_policy(domain) values('" +
urlName + "')");
PreparedStatement pstmt=con.prepareStatement(query);
pstmt.executeUpdate(query);
pstmt.close();
} else {
System.out.println(" - none");
String query =("insert into P3Poldsubsub1.no_policy(domain) values('" + urlName
+ "')");
PreparedStatement pstmt=con.prepareStatement(query);
pstmt.executeUpdate(query);
```

```
pstmt.close();
}
}
catch (Exception e) {
if(urlName != null) {
String query =("insert into P3Poldsubsub1.bad_domain(domain) values('" +
urlName + "')");
PreparedStatement pstmt=con.prepareStatement(query);
pstmt.executeUpdate(query);
pstmt.close();
}
System.out.println(e.getMessage());
}
finally
{
link.disconnect();
}
}
}
} catch (Exception e) {
System.out.println(e.getMessage());
```

```
}

}

}
```

Appendix-10

iWatch Database Schema

```
DROP TABLE IF EXISTS `tbl_domains`;

CREATE TABLE `tbl_domains` (

  `domain` varchar(255) NOT NULL,

  `number` int(11) unsigned NOT NULL default '0',

  PRIMARY KEY  (`domain`),

  UNIQUE KEY `name` (`domain`),

  KEY `tbl_domains_idx` (`domain`),

  KEY `domain_idx` (`domain`),

  KEY `number_idx` (`number`)

) ENGINE=MyISAM DEFAULT CHARSET=latin1;



DROP TABLE IF EXISTS `tbl_events`;

CREATE TABLE `tbl_events` (

`date` timestamp NOT NULL default CURRENT_TIMESTAMP on update
  CURRENT_TIMESTAMP,

  `event_type` varchar(100) default NULL,

  `event_description` longblob,

  `comment` blob,

  `id` bigint(22) NOT NULL auto_increment,

  `id_url` longblob,
```

```
  PRIMARY KEY  (`id`)
) ENGINE=MyISAM DEFAULT CHARSET=latin1;


DROP TABLE IF EXISTS `tbl_global`;
CREATE TABLE `tbl_global` (
  `name` varchar(80) NOT NULL default '',
  `value` bigint(22) default NULL,
  PRIMARY KEY  (`name`)
) ENGINE=MyISAM DEFAULT CHARSET=latin1;
DROP TABLE IF EXISTS `tbl_seedlist`;
CREATE TABLE `tbl_seedlist` (
  `count` int(11) default NULL,
  `isNew` tinyint(1) default NULL,
  `link_from` bigint(20) default NULL,
  `id` int(11) NOT NULL auto_increment,
  `name` varchar(255) NOT NULL default '',
  PRIMARY KEY  (`name`),
  UNIQUE KEY `name` (`name`),
  KEY `id` (`id`)
) ENGINE=MyISAM DEFAULT CHARSET=latin1;
```

```
DROP TABLE IF EXISTS `tbl_sitestack`;

CREATE TABLE `tbl_sitestack` (

  `id` bigint(22) unsigned NOT NULL auto_increment,

  `today_date` timestamp NOT NULL default CURRENT_TIMESTAMP on update
   CURRENT_TIMESTAMP,

  `name` varchar(255) NOT NULL default '',

  `domain` varchar(100) NOT NULL default '',

  `link_from` longblob,

  PRIMARY KEY  (`id`),

  KEY `id` (`name`),

  KEY `tbl_sitestack_idx` (`id`),

  KEY `domain_idx` (`domain`),

  KEY `idx` (`id`),

  KEY `namex` (`name`),

  KEY `domainx` (`domain`)

) ENGINE=MyISAM DEFAULT CHARSET=latin1;


DROP TABLE IF EXISTS `tbl_urlhistory`;

CREATE TABLE `tbl_urlhistory` (

  `name` varchar(255) NOT NULL default '',

  `thread_id` bigint(20) default NULL,

  `use_form` tinyint(4) default NULL,
```

```
`use_cookies` tinyint(4) default NULL,

`use_p3p` tinyint(4) default NULL,

`use_webbug` tinyint(4) default NULL,

`today_date` timestamp NOT NULL default CURRENT_TIMESTAMP on update
CURRENT_TIMESTAMP,

`domain` varchar(90) default NULL,

`js_annoyances` bigint(20) default NULL,

`js_events` bigint(20) default NULL,

`html_annoyances` bigint(20) default NULL,

`content_cookies` bigint(20) default NULL,

`refresh_tags` bigint(20) default NULL,

`unsolicited_popups` bigint(20) default NULL,

`all_popups` bigint(20) default NULL,

`img_reorder` bigint(20) default NULL,

`banners_by_size` bigint(20) default NULL,

`banners_by_link` bigint(20) default NULL,

`tiny_textforms` bigint(20) default NULL,

`jumping_windows` bigint(20) default NULL,

`frameset_borders` bigint(20) default NULL,

`demoronizer` bigint(20) default NULL,

`shockwave_flash` bigint(20) default NULL,

`quicktime_kioskmode` bigint(20) default NULL,
```

```
  `site_specifics` bigint(20) default NULL,
  `count` int(11) unsigned NOT NULL default '0',
  `id` bigint(22) unsigned NOT NULL auto_increment,
  `IpAddress` varchar(255) default NULL,
  PRIMARY KEY  (`name`),
  UNIQUE KEY `name` (`name`),
  KEY `id` (`id`),
  KEY `tbl_urlhistory_idx` (`id`),
  KEY `tbl_sitestack_idx` (`id`),
  KEY `tbl_id_urlx_namex` (`name`),
  KEY `domain_idxx` (`domain`),
  KEY `name_idxa` (`name`),
  KEY `namex` (`name`),
  KEY `banner_idx` (`banners_by_size`)
) ENGINE=MyISAM DEFAULT CHARSET=latin1;
```