



## APPROVAL SHEET

Title of Dissertation: Data Mining Approach to Compare Privacy Policies

Name of Candidate: Shaikha Alduaij  
Doctor of Philosophy, 2017

Dissertation and Abstract Approved:

Dr. Zhiyuan Chen  
Associate Professor  
Information Systems Department

Date Approved: \_\_\_\_\_

## ABSTRACT

Title of thesis: A DATA MINING APPROACH TO COMPARE  
PRIVACY POLICIES

Shaikha Alduaij, Doctor of Philosophy, 2017

Directed By: Dr. Zhiyuan Chen, Associate Professor,  
Information Systems Department, UMBC

As it becomes easier and less expensive for service providers to store huge amounts of data, the information collected about individuals is growing rapidly. At the same time, individuals' concerns about their privacy is increasing. Although most service providers use privacy policies to explain what information they are collecting, who will access it, and for what purpose, existing research shows that users often do not read privacy policies or they find privacy policies difficult to understand. Thus, users may not make the right regarding securing their privacy. Some studies have proposed tools to enhance the effectiveness of privacy policies and thus facilitate decision making, whereas others have introduced visualization models to increase privacy usability and effectiveness. Despite all of this, there has been relatively little work done on providing a comparison model to assist users when comparing the privacy practices of different companies in an effort to make informed decisions. In this study, we first analyze users' awareness of privacy policies and the privacy practices described in them, their privacy concerns, and their privacy needs. Next, we use text mining techniques to extract information users care about such as collected information, shared information, and provided

controls. Unlike existing techniques, our approach attempts to avoid the use of patterns or rules as much as possible because the format of privacy policies often changes over time, and therefore, patterns and rules often become obsolete. We then develop a comparison tool to show the extracted information side by side. Then we conduct a survey to validate our comparison tool and gather users' privacy preferences. Because a side-by-side comparison may not work well when users are comparing a large number of policies, we propose a data mining based method to rank privacy policies. Unlike existing techniques that rely on user ratings, which are often not reliable, our approach relies on pair-wise preferences given by users, which are often a lot more reliable.

DATA MINING APPROACH TO COMPARE PRIVACY POLICIES

By

Shaikha Alduaij

Dissertation submitted to the Faculty of the Graduate School of the

University of Maryland, Baltimore County, in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

2017

© Copyright by  
Shaikha Alduaij  
2017



## Table of Contents

List of Figures.....	iv
List of Tables.....	iv
CHAPTER 1: Introduction .....	1
1.1 Motivation.....	1
1.2 Proposed work.....	3
1.3 Research contribution.....	5
1.4 Organization of Dissertation Thesis .....	6
CHAPTER 2: Literature Review .....	7
2.1 Limitations of privacy policies.....	7
2.2 Privacy Concerns .....	10
2.3 Factors affecting users’ privacy concerns .....	12
2.4 Risks of information breaches.....	13
2.5 Privacy policy tools.....	16
2.6 Privacy policy visualization.....	20
CHAPTER 3: Analysis of User’s Privacy Concerns .....	25
3.1 Introduction .....	25
3.2 Breaking down privacy policies.....	26
3.3 Research questions .....	28
3.4 Study design.....	29
3.5 Results.....	33
1. Does users’ demographic information affect their concerns? .....	33
2. Do users’ frequency of usage and awareness of the service privacy policy affect their privacy concerns? .....	34
3. Do users prefer more controls? .....	35
4. Are users aware of the provided controls? Do they use them? .....	36
5. Which data flow or data type is the most unexpected by users, and which data flow or data type are users most uncomfortable with?.....	37
6. Do users have trouble identifying the reason service providers collect and share information? ..	39
7. Does clarifying the purpose of the information collection or sharing ease users’ concerns about the privacy of their data? .....	41

8. Is it possible to predict the level of privacy concerns for a specific user? .....	42
9. What other factors affect users' privacy concerns? .....	44
3.6 Some guidelines .....	45
3.7 Discussion and Conclusion .....	47
CHAPTER 4: Information Extraction .....	50
4.1 Introduction .....	50
4.2 Information extraction approach.....	51
4.2.1 Data collection .....	51
4.2.2 Preprocessing.....	52
4.2.3 Personal information .....	53
4.2.4 Controls extraction.....	60
4.2.5 Sharing extraction .....	63
4.3 Results and Evaluation .....	67
4.3.1 Personal information evaluation .....	67
4.3.2 Controls extraction evaluation.....	75
4.3.3 Sharing extraction evaluation .....	78
4.4 Discussion and conclusion .....	79
CHAPTER 5: Privacy Policies Comparison.....	82
5.1 Introduction .....	82
5.2 Privacy database .....	83
5.2.1 Information category .....	83
5.2.2 Database schema .....	84
5.3 Comparison tables .....	85
5.4 Research questions .....	87
5.5 Study design .....	87
5.5.1 Survey sample .....	90
5.7 Results .....	95
5.7.1 Timing results.....	95
5.7.2 Accuracy.....	98
5.7.3 Preferences .....	102
CHAPTER 6: Ranking Privacy Policies .....	103

6.1 Overview .....	103
6.1.1 Overview of pairwise comparison .....	105
6.2 Data collection and labeling.....	106
6.3 Learn to rank .....	108
6.4 Feature selection .....	109
6.4.1 Generalized Feature Selection .....	110
6.4.2 Counting Feature Selection.....	111
6.4.3 Minimum redundancy feature selection .....	111
6.5 Classification methods .....	112
6.6 Ranking Validation .....	113
6.7 Discussion and conclusion .....	118
CHAPTER 7: Conclusion.....	119
References .....	122

## List of Figures

Figure 1: P3P Expandable Grid.....	21
Figure 2: Nutrition Label .....	22
Figure 3: Layered privacy policy.....	24
Figure 4: Data flow graph for Bank of America.....	28
Figure 5: Sample questions.....	31
Figure 6: Average comfort level based on age .....	33
Figure 7: Average comfort based on gender .....	33
Figure 8: Average comfort based on frequency of usage .....	35
Figure 9: Average comfort level based on awareness of the privacy policy.....	35
Figure 10: Tagged sentence .....	65
Figure 11 Overall steps of privacy policies comparison.....	83
Figure 12: Schema design .....	85
Figure 13: An example of binary labelling.....	107
Figure 14: An example of proportional labelling .....	108
Figure 15: Learning to rank.....	109
Figure 16: An example of the SMOreg classifier model.....	115

## List of Tables

Table 1: Policy segments from Bank of America.....	27
Table 2: Age distribution of the participants .....	32
Table 3: Gender distribution of the participants .....	32
Table 4: Percentage of control preferences .....	35
Table 5: Percentage of participants' awareness of and usage of Facebook existing controls .....	36
Table 6: Percentage of participants' awareness of and usage of Twitter existing controls .....	36
Table 7: Percentage of participants' awareness of and usage of Bank of America existing controls .....	36
Table 8: Percentage of expectation and average comfort level based on type of flow .....	38
Table 9: Percentage of expectation and average comfort level based on type of data.....	39
Table 10: Type of data .....	40
Table 11: Type of flow.....	41
Table 12: Comparison of comfort ratings between the expectation condition (2 <sup>nd</sup> column) and the purpose condition (3 <sup>rd</sup> column) based on data type. ....	42
Table 13: Comparison of comfort ratings between the expectation condition (2 <sup>nd</sup> column) and the purpose condition (3 <sup>rd</sup> column) based on type of flow .....	42
Table 14: Variables type.....	43
Table 15: R <sup>2</sup> values for the results of logistic regression modelling .....	43
Table 16: R <sup>2</sup> values for the results of linear regression modelling .....	44
Table 17: Independent attributes that have significant correlation with comfort levels.....	45

Table 18: List of domains and service providers.....	52
Table 19: An example of rules table .....	57
Table 20: List of action words .....	58
Table 21: List of control keywords .....	61
Table 22: Control related terms.....	62
Table 23: Sample of analysis results .....	66
Table 24: Random Forest evaluation results using different dimensions .....	68
Table 25: Sampling evaluation results .....	69
Table 26: Testing dataset evaluation results- personal information rule-based extraction .....	70
Table 27: Results of all output – no action words .....	73
Table 28: Results of distinct outputs – no action words.....	73
Table 29: Results of distinct outputs – action words.....	73
Table 30: Results of all output – action words.....	73
Table 31: Training dataset evaluation results- personal information matching-based extraction .....	74
Table 32: Results of distinct outputs – no action words.....	74
Table 33: Results of distinct outputs – action words.....	74
Table 34: Results of all outputs – no action words.....	74
Table 35: Results of all outputs – action words .....	74
Table 36: Testing dataset evaluation results- personal information matching-based extraction.....	75
Table 37: Results of distinct outputs – no control keywords .....	76
Table 38: Results of all outputs – no control keywords.....	76
Table 39: Results of distinct outputs – control keywords.....	76
Table 40: Results of all outputs – control keywords.....	76
Table 41: Training dataset evaluation results- controls extraction .....	76
Table 42: Results of distinct outputs – no control keywords .....	77
Table 43: Results of all outputs – no control keywords.....	77
Table 44: Results of distinct outputs – control keywords.....	77
Table 45: Results of all outputs – control keywords.....	77
Table 46: Testing dataset evaluation results- controls extraction.....	78
Table 47: Training dataset evaluation results- sharing extraction .....	78
Table 48: Testing dataset evaluation results- sharing extraction .....	79
Table 49: Categorization of privacy practices .....	84
Table 50: Personal information comparison table .....	85
Table 51: Controls comparison table .....	86
Table 52: Sharing comparison table .....	86
Table 53: Service providers anonymized name .....	88
Table 54: Age distribution of the survey participants.....	94
Table 55: Average time for health domain .....	96
Table 56: Average time for financial domain.....	96
Table 57: Average time for shopping domain.....	96
Table 58: Average time for networking domain .....	96

Table 59: Average time spent by the participant each domain and each version and the aggregated average.....	98
Table 60: The average accuracy of participants getting correct answer about specific information and about the company who collects more data of health domain. ....	99
Table 61: The average accuracy of participants getting correct answer about specific information and about the company who collects more data of financial domain. ....	99
Table 62: The average accuracy of participants getting correct answer about specific information and about the company who collects more data of shopping domain. ....	99
Table 63: The average accuracy of participants getting correct answer about specific information and about the company who collects more data of social networking domain. ....	100
Table 64: P-value of domains across all survey versions. ....	101
Table 65: The accuracy percentage of participants' responses across different versions of the survey. ....	101
Table 66: Participants' preferences for list and table versions of all domains. ....	102
Table 67: Pairwise comparison of the health domain ....	105
Table 68: Pairwise comparison of the financial domain.....	105
Table 69: Pairwise comparison of the shopping domain.....	106
Table 70: Pairwise comparison of the networking domain.....	106
Table 71: R <sup>2</sup> of SMO regression classification.....	114
Table 72: R <sup>2</sup> of linear/logistic regression classification.....	115
Table 73: Health domain comparison results.....	117
Table 74: Financial domain comparison results.....	117
Table 75: Shopping domain comparison results.....	117
Table 76: Social networking comparison results.....	117

## CHAPTER 1: Introduction

The use of online services has become a necessity because it is involved in most individuals' daily activities, including those related to business, education, and communication. When using these services, users usually share their information for purposes such as registering a service, customizing their experience, or sharing their thoughts and interests with others. The collection and storage of this vast amount of information has raised users' concerns about how these practices will affect their information privacy. Natural questions arise in this context, such as how individuals' information will be stored, who will access it, how it will be used, and for what purpose.

To allay users' concerns, most of the service providers explain their practices with privacy policies. A privacy policy is a statement or legal document provided by the service provider to explain the handling of the user's gathered information by describing what information will be collected, how it will be used, with whom it will be shared, and the purpose of that sharing. Furthermore, it describes users' rights and options to change some of the practices.

### 1.1 Motivation

It is important that a privacy policy is written and presented to users clearly so they can understand how their information is being used. For example, Thelma Arnold, a user of the AOL search engine, did not know that AOL stored and shared her queries. Her identity was disclosed

to the public by inferring her identity based on her queries. She stated in an interview, “My goodness, it’s my whole personal life ... I had no idea somebody was looking over my shoulder” (Barbaro & Zeller, 2006). Her unawareness of this practice could be because it was not clearly described in the AOL privacy policy or because she did not read the policy.

There are several limitations associated with the current privacy policies. First, a privacy policy can be difficult to understand. Numerous studies have stated that privacy policies usually use difficult linguistic terms to describe their practices and require a high level of education to be understood (Jensen & Potts, 2004; Pollach, 2007; Sumeeth, Singh, & Miller, 2010). A study by McDonald et al. (2009) stated that users find it hard to correctly answer questions related to privacy practices when they are provided with the policies. Additionally, the extensive time required to read a privacy policy often leads people to avoid reading them. McDonald and Cranor (2008) estimated that it would take approximately 244 hours per year for users to read every unique service provider’s privacy policy. Other studies have found that several privacy policies poorly cover the practices that raise users’ privacy concerns regarding their data collection, data storage, and data sharing (Earp, Antón, Aiman-Smith, & Stufflebeam, 2005; Pollach, 2007). Finally, the platform for privacy preferences (P3P) and trustable seals proposed as improvements of current privacy policies forms has tended to overcome current policies’ limitations, but they have suffered from low adoption due to difficulties of implementation.

Improving privacy policies or providing tools to enhance them is important. Making them clear and easy to understand would increase their readability and thus users’ awareness of privacy practices. Some users accept the terms and conditions before using a service, and

they are not aware of the risks that may be associated with the sharing of their personal data. The information users post or share online may be misused or used for purposes they are not aware of intentionally or accidentally. The risks or negative effects of using individual personal information can lead to embarrassment, decreased opportunity of employment, identity theft, cyber stalking, or phishing. These risks can happen by using information users themselves share without knowing how badly it can be misused. For example, some users may share on social media a photo for a ticket of an event they plan to attend. Any person who can view the picture may copy the barcode, print it, and use it (Ehling, 2013). Risks can also happen by using information stored in servers or shared with a third party. For example, an individual's identity may be associated with a disease when the information he searched for in a website is shared with third parties, disclosed to the public, or given to a data broker. This might affect users by loss of employment (Libert, 2015; Walters & Betz, 2012). Indeed, the availability of an individual's information can even lead to murder. Amy Boyer, a 20-year-old women from Nashua, was murdered after the criminal stalked her and gathered information about her work location using online sources (Donovan & Bernier, 2008).

## 1.2 Proposed work

Several research studies have tried to propose different solutions to improve privacy policies and increase users' privacy awareness. Some have introduced tools that help in privacy policy development and enforcement; others have introduced standardized presentations of privacy policies.

Maintaining users' privacy directly affects a company's reputation and customers' satisfaction (Mont, 2004; Mont, Thyne, & Bramhall, 2005). Many studies have examined tools aimed to help policy authors to develop and enforce their policies, such as the Server Privacy Architecture and Capability Enablement, Privacy Policy Modeling Language Processor (Brodie et al., 2005; Yu & Murthy, 2007), and other approaches directed to help privacy enforcement of policies for authors and users. The Platform for Privacy Preferences and User Privacy Policy are two examples of these approaches (Aïmeur et al., 2009; Tsai, Egelman, & Cranor, 2011). These solutions require the privacy policy to be written in a machine-readable language which makes it difficult to implement.

Other approaches avoid this limitation and use the existing natural language policies to develop the tools. These approaches' main goal is to present evaluation of the service provider's privacy practices and assist the user to judge how good the company is in terms of privacy practices (Costante, Hartog, & Pekovie, 2013; Costante & Sun, 2012; Vimercati et al., 2009).

Visualizing and standardizing privacy policies are parts of the proposed solutions to improve their effectiveness. P3P Expandable Grid, "nutrition label," and layered privacy notices are different forms of privacy policies, which tend to present privacy practices in a simplified display aimed to improve users' awareness about these practices.

Although visualization and standardizing, privacy policies facilitate privacy comparisons by showing information consistently, to our best knowledge none of the solutions tend to allow users to compare policies by showing them side by side. In this study, we will present a solution

that aims to help users compare privacy policies to make informed decisions and choose the company that meets their privacy needs.

### 1.3 Research contribution

Our thesis aims to facilitate the comparison of service providers' privacy practices by analyzing their privacy policies, highlighting the policies' main components, asking users about their preferences, and then ranking the policies based on their preferences. Specifically, the research contributions of our thesis can be summarized as follows:

- Investigate users' awareness, concerns, and needs regarding information privacy. We conducted a survey that examines users' familiarity with some privacy practices, how users are concerned regarding these practices, and the factors that affect their concerns.
- Analyze privacy policies from several domains and extract information related to practices users care about and propose extraction algorithms to extract information from privacy policies. Unlike existing techniques, our approach tries to avoid the use of patterns or rules as much as possible because the format of privacy policies often changes over time and patterns and rules often become obsolete. Instead, our approach is based on matching popular and common keywords.
- Propose a privacy comparison model which shows extracted information side by side. We conducted a survey to examine the effectiveness of the comparison solution and to gather users' privacy preferences based on services providers' privacy practices.

- Propose a scoring function to assign scores to privacy policies making it possible to rank them. Side-by-side comparison does not work well when users are comparing many policies. We propose a data mining based method to rank privacy policies. Unlike existing techniques that rely on user ratings, which are often not reliable, our approach relies on pairwise preference given by users, which is often a lot more reliable.

#### 1.4 Organization of Dissertation Thesis

The remainder of the dissertation thesis is organized as follows. We will present a comprehensive literature review on privacy policies in Chapter 2.

Chapter 3 will give a detailed description of the analysis of user's privacy concerns. In Chapter 4, we will present the proposed information extraction algorithms. Then we will describe the comparison model and comparison study design in Chapter 5. In Chapter 6 we will generally discuss the privacy policies ranking methodology.

## CHAPTER 2: Literature Review

### 2.1 Limitations of privacy policies

Although privacy policies are meant to explain to Internet users everything related to the collection and sharing of their personal information when using the service of any individual provider, existing studies have found that these policies have many negative issues. To begin, most privacy policies are not easily accessible, comprehensive, or manageable for the average user. They often fail to meet users' actual privacy concerns, and improvements have been arguably few and quite limited.

First, privacy policies are quite difficult to understand. McDonald et al. (2009) performed a comparative study of the formats of six different companies' privacy policies and found that users considered every studied format difficult to understand. Pollach (2007) discovered that privacy policies are written in unclear and misleading linguistic patterns. Additional studies have found that privacy policies require a high level of education to be fully understood. Jensen and Potts (2004) found that people without a high school education can only understand 6% of privacy policies, and according to Sumeeth, Singh, and Miller (2010), 20% of the policies they analyzed require a postgraduate education to be understood. Moreover, privacy policies often contain long sentences and domain-specific terms unfamiliar to typical users, making them difficult to read and comprehend (Jensen & Potts, 2004). Additionally, policies are often written in an unclear and confusing manner (McDonald et al.,

2009). For example, a company states in its online policy, “While Company does not currently support telemarketing, it is possible that in the future Company properties may contact you by voice telephone,” to explain that that they may use information for telemarketing.

The second issue is that it takes a lot of time to read privacy policies. McDonald and Cranor (2008) indicated that privacy policies are often quite long. The average length is about 2,500 words, which is equivalent to a paper of five to six pages written in a 12-point font. The authors asked, “If website users were to read the privacy policy for each site they visit just once a year, what would their time be worth?” They calculated the average time of reading privacy policies in two ways. First, they analyzed the word count of the privacy policies for the 75 most frequently visited websites according to AOL search data. Second, they conducted a survey and calculated the average time spent by 212 participants to skim these privacy policies. Their results showed that it would take an individual 244 hours per year to read and 154 hours per year to skim every unique privacy policy for websites visited. Additionally, by estimating the value of time as 25% of average hourly salary for leisure and twice wages for time at work, the results indicate that the national cost to read these privacy policies would be \$781 billion per year, and the cost to skim these policies would be \$492 billion per year. Jensen and Potts (2004) added that the results of a survey done in a university setting for a stand-alone website requiring registration showed that due to its length very few participants visited the privacy policy page, specifically, only in 0.24% of 55,158 sessions was the policy viewed.

The third issue is that privacy policies do not consider consumers’ common concerns about their information collection, transfer, and storage process and overall service provider’s

practices. For example, Earp et al. (2005) found that privacy policies of 50 websites did not emphasize issues commonly raised in Internet users' privacy concerns such as revealing information about their identity or information related to their activities. Pollach (2007) also found that privacy policies poorly cover users' privacy concerns regarding data collection, data storage, data sharing, and unsolicited marketing communications. He conducted a survey to verify the coverage of these concerns in the privacy policies and found that 39.4% of the questions about these practices could not be answered due to the lack of sufficient information provided by the examined policies.

The fourth issue is low adoption of proposed solutions. Several studies have introduced approaches and tools for privacy policy improvements. For example, P3P is a platform for privacy preferences introduced by the World Wide Web Consortium (W3C) to enable browsers to read a website's privacy policy and match it to users' predefined preferences. However, a large number of websites had errors when they adapted the P3P policy format (Leon, Cranor, McDonald, & McGuire, 2010). Privacy trustable seals is another suggested improvement for privacy policies. This program reviews and evaluates a service provider's policy, matches it to the website's privacy practices, and displays a trust mark on the website informing users of the evaluated privacy practices. This helps users save time by only requiring them to verify the seals displayed on the website instead of reading the full policy. Nevertheless, the results of Moore's (2005) study showed that users found it hard to recognize the graphic seals and distinguish between actual and fake seals, concluding that consumers may simply be responding to a graphic design, rather than to any attempt by the website to be certified as trustworthy.

Additionally, only 34.3% of Pollach's (2007) survey questions relating to data handling of providers with at least one trust seal were answered correctly, which indicates that the coverage of privacy issues on privacy policies is another limitation of privacy trustable seals.

Our study tries to alleviate some of these limitations, including the difficulty of understanding a policy and the long time needed to read a policy.

## 2.2 Privacy Concerns

Smith, Dinev, and Xu (2011) found that the efficiency of information storage and usage by service providers raised privacy concerns. Dinev and Hart (2004) showed that privacy concerns and a low willingness to provide personal information are caused by the conflict of the need for information to be collected and the resulting threat to users' privacy. Westin (2001) reported that 90% of Americans are worried that their personal information will be misused, and specifically 77% describe themselves as "very concerned."

Researchers have different findings when it comes to privacy controls. The survey *Consumer action "Do not track"* (2013) results showed that 87% of their participants strongly agreed that they have the right to make choices regarding their information privacy and want more control. According to Hazari & Brown (2013), the employed users of social networks websites are more likely to control their privacy preferences than the unemployed users because it can have a negative influence on their employment status and relationships. Young and Quan-Haase (2009) found that students made some effort to control their personal information in their profiles to protect themselves. For example, they provide limited

information in their profiles to prevent other users from gaining comprehensive information about them. Moreover, Harris Interactive (2001) reported that almost 80% of Americans think that it is “very important” to control collected information. Although existing research showed that people need control over their information privacy and think this control is important, other research found that people provide much information without using the available controls or reading the privacy statements. Gross and Acquisti (2005) found that only 1.2% and 0.06% of their study participants changed default profile searchability and visibility, respectively. Hazari and Brown (2013) stated that users post a great deal of content to customize their experience using different services in a way that negatively affected their privacy. Further, some studies have shown that users do not read privacy notices or policies provided by the web services that they use (Chaianuchittrakul, 2013; Lin et al., 2012). In this study, we will investigate the provision of controls for each privacy policy segment, whether users make use of the provided controls, and whether using these controls lowers users’ privacy concerns.

Although existing studies have specified users’ privacy concerns, there has been relatively little work done on how consumers respond to the policies’ individual segments such as sharing profile information with a third party for marketing purposes. In this study, we use a crowd sourcing approach to analyze consumer responses to individual privacy policy segments. We overcome the readability issue of privacy policies by conducting the study at a micro level (i.e., segment level) because it is much easier for consumers to understand an individual policy

segment than the full text of the policy. We also parse each privacy policy segment into easy-to-understand components.

### 2.3 Factors affecting users' privacy concerns

There are multiple factors that affect an individual's privacy concerns and practices. Lin et al. (2012) found that users of mobile applications feel more comfortable and less concerned when they are notified of the reasons for which their information is collected by the application. For example, participants felt less concerned when they were informed that the application Dictionary collected their location for providing words trending in their area.

The impact of demographic factors such as age and gender on individual concerns has been the focus of several studies with some different findings. In a study about the impact of gender on online privacy concerns for undergraduate students at a Southwestern American university, Yao et al. (2007) found that online concerns did not vary between gender groups. Nowak and Phelps's (1992) results also supported this finding, stating that users' concerns about threats to personal privacy did not correlate with gender. However, Janda and Fair (2004) found in their study of non-student adult Internet users that gender had a significant impact on privacy concerns showing that women have greater privacy concerns including fraud, privacy, security, hacking, and child protection. Their results also strongly indicated that older adults are more concerned about their online privacy. Tufekci (2008) stated that both age and gender have an impact on undergraduate student concerns and disclosure behavior when using social networking sites. The results showed that younger users tended to display more information

and that there were significant gender differences based on the information displayed. For example, men tended to display their phone numbers in their account profiles more than women, while women tended to disclose information related to their religion and personal preferences such as favorite music and books more than men.

In our survey, we analyzed factors affecting users' privacy concerns including demographics information, the service provider's service nature, and the type of information.

#### 2.4 Risks of information breaches

In this digitalized age people are using online services daily. They may use it for education purposes, travel reservations, social networking, doctors' appointments managements, and many other daily activities. Using these services involves the gathering of users' personal information including identity, financial, and health information. The availability of this valuable information improves users' vulnerability of privacy threads. Following are some examples of the risks users may face when their information is incidentally or intentionally disclosed.

Violation of any kind of private information could easily bring embarrassment. Publishing personal information relating to specific activities can even lead to more serious circumstances. The Ashley Madison hack captured public attention recently. A group of hackers stole personal information from the adult dating website, Ashley Madison, and they disclosed more than 30 million users' information including full names, street addresses, and e-mail addresses. Although the company announced that the stolen data do not contain financial

information, subscribers were concerned that their reputations, marriages, jobs, and lives could be at stake. The case got even worse when two suicides were linked to that information leak (Mansfield-Devine, 2015).

Publishing personal information by an individual, especially on social network websites, can decrease her opportunity to get a job (CareerBuilder.com, 2009). The results of the CareerBuilder survey reported that 45% of employers used social networking websites to assess job candidates. The results showed that employers eliminated the job's candidates for some of the following reasons:

1. Candidate posted provocative or inappropriate photographs or information.
2. Candidate posted content about their drinking or using drugs.
3. Candidate bad-mouthed their previous employer, coworkers, or clients.
4. Candidate showed poor communication skills.
5. Candidate made discriminatory comments.
6. Candidate lied about qualifications.

The availability of an individual's data in several websites servers could make her information vulnerable to identity theft. Gaining access to one source of information and linking it to other sources would help to get sufficient data about an individual. Identity theft involves gaining financial or medical advantages by using another person's name or identity. A survey report of identity theft in the United States showed that in 2005 consumers lost about \$56 billion because of companies' data breaches (Javelin Strategy & Research, 2006). In

addition to financial identity theft, a report by the Federal Trade Commission (2010) stated that from 2001–2006 more than 250,000 identity theft cases counted as medical identity theft. The victim of medical identity theft may be negatively affected by the falsifying of his medical history, which may lead to inappropriate medical diagnoses, loss of employment, and emotional consequences (Betz, 2012; Cullen, 2007; Federal Trade Commission, 2010; Identity Theft Resource Center, 2007; Schmidt & McCoy, 2008).

Cyber stalking is another issue related to data breaches. The use of personal information for stalking an individual can lead to serious consequences, including murder. The murder of Asia McGowan is an example of a cyber-stalking crime. The media reported that the killer, Anthony Powell, stalked the victim on Facebook and YouTube before he killed her (Reisman, 2009).

Phishing is a method to gain the victim's personal data usernames, passwords, and credit card details in the form of a "SPAM" attack. The information is usually obtained from users through electronic communication. An attacker with previous knowledge about an individual can easily gain additional information and use it for malicious reasons (Reilly, 2006). A survey of 1,335 US.net users reported that in 2004 US consumers lost about \$500 million because of phishing scam attacks (Leyden, 2004).

Finally, users' health interests or health searches can be used by third parties and have a negative impact. First, it may lead to personal identification by associating an individual's name to a disease. This can happen when a person searches for a disease, symptom, or treatment on a website that shares user's queries with a third party. An attacker getting access to this

information, incident leakage, or the selling of information by brokers can reveal the association between the user and a disease to the public. Second, this can lead to blind discrimination where the user's identity is not necessarily disclosed, but she may be treated differently. For example, marketing companies may exclude an individual from receiving some offers based on perceived medical conditions (Libert, 2015).

Our study can help users understand and compare policies and may lead users to make intelligent privacy choices to lower their privacy risks.

## 2.5 Privacy policy tools

Privacy management is important for enterprises and organizations that handle the identity and personal data of customers, employees, and business partners. It has implications on their compliance with regulations, their reputation, and customers' satisfaction (Casassa, 2004; Casassa et al., 2005). Studies have introduced several approaches supporting privacy enforcement, including SPARCLE and PPMLP.

Server Privacy Architecture and Capability Enablement (SPARCLE) is a tool developed to assist privacy policies' authors during the process of privacy policy formulating, auditing, and enforcing. The main concept of this approach is to apply natural language processing techniques on policies written in constrained natural language. It parses the privacy rules written in specific patterns to identify policy elements and generate machine readable language such as EPAL to facilitate the creation and management of the policies (Brodie et al., 2005).

Privacy Policy Modeling Language Processor (PPMLP) is another solution to assist policies authors to generate privacy policies that match the organization's privacy principles. It requires the authors to identify the privacy policy specifications. The specifications are mapped to predefined meta-privacy policy specifications, and then it parses the user's input based on a set of privacy policy grammar rules and translates the natural language policy into corresponding machine readable language (EPAL). The system then checks the compliance of the privacy policy with the privacy policy principles (Yu & Murthy, 2007).

Also, Joshi et al. proposed a semantic approach to automate the management of Big Data privacy policies. They analyzed the privacy policy of ten cloud providers, provided a semantic ontology to describe the privacy policies text. They used NLP and text mining techniques to extract rules from these policies. Then, they suggested to use the ontology and extracted rules to automate the creation and validation of privacy policies documents for cloud providers (Joshi et al, 2016)

Other proposed approaches aim to assist privacy enforcements directed to end-users and privacy authors including P3P and UPP. Platform for Privacy Preferences (P3P) is an architecture designed to inform users about websites' data collection practices. It requires the website to be written in machine-readable language, XML, to evaluate the privacy contents and verify that they match predefined users' privacy preferences. Users do not need to read the full policy, only the P3P. There are different implementations of P3P including Privacy Bird, a browser extension that alerts users about privacy practices of a visited website based on their predefined preferences (Carnor, Arjula, & Guduru, 2002), and Privacy Finder, a P3P enabled

search engine that only returns search results that meet users' privacy preferences (Tsai, Egelman, & Cranor, 2011).

User Privacy Policy (UPP) is a domain-specific approach to enforce privacy practices of social networks services. It allows users of social networks to define their privacy preferences and then communicate these preferences to websites before allowing access to their data by other users, service providers, or third parties (Aïmeur et al., 2009).

One of the major limitations of these approaches is that they need server-side implementation. This fact leads to low adoption of these solutions because service providers must implement their policies in machine-readable language in addition to a natural language version. Beatty et al. (2007) showed in their study results that only 20% of 300 top e-commerce websites are P3P enabled.

Instead of creating a machine-readable language version of a privacy policy, other studies have proposed policy enforcement approaches aiming to inform end users about privacy practices with the use of existing natural text policies. The PrimeLife Privacy Dashboard is a browser extension that helps to inform users about privacy practices and the kind of information collected by websites they visit. For example, it notifies users if a visited website collects cookies, has a P3P policy, and is certified by a trust seal. The tool allows users to determine privacy preferences for a website and set permissions for data access. In addition, it adopts the smiley face icons to rate websites based on the websites' privacy practices such as cookies collection, use of an external third party, and implementation of P3P policy. One remarkable limitation of this tool is that one of the rating criteria is the existence of P3P policy,

which we mentioned earlier for its low adoption, which lowers most websites' ratings (Vimercati et al., 2009).

The Terms of Service; Didn't Read project is an example of studies that aim to notify users of websites' privacy practices. It is a web browser extension that displays an icon that classifies the visited website based on pre-evaluated privacy topics and contents. The evaluation of the topics and contents is achieved using Crowdworkers, which reviews and evaluates individual topics of the policy by assigning scores and badges. Once a service has enough badges to assess the fairness of their topics for users, a class is assigned automatically by calculating the average scores. The main limitation of this solution is that it depends on Crowdworkers to rate policies, so only well-known websites will be rated, and less visited websites may be ignored.

Zimmeck and Bellovin (2014) used the same concept with an improvement of adding an automatic classifier of privacy topics for the companies not evaluated in the pre-existing crowd sourcing repository. Their classifier is based on a binary classifier that determines the existence of specific terms.

Costante and Sun (2012) proposed a solution to automatically analyze privacy policies and evaluate their completeness. They defined a set of privacy categories using text categorization and machine learning techniques to analyze the natural language privacy policy and assess whether it covers these categories or not. Examples of privacy categories included in the evaluation are advertising, collection, location, cookies, and policy change. Once the

covered categories are determined, a grade is assigned to the policy representing the policy's privacy protection level.

Costante, Hartog, and Pekovie (2012) proposed parsing privacy policies and extracting what information is collected by these policies. They applied information extraction techniques to the task of extracting structured text from unstructured or semi-structured documents to extract a list of the personal information collected by a certain company based on the content of its policy. The system consists of several modules including a part-of-speech (POS) tagger that annotates words with a related POS tag and name entity (NE) recognition module that seeks to identify and classify elements in texts into predefined categories. Their idea is to develop a browser extension that presents extracted information to the end users, informing them about policy contents and to allow them to make informed privacy decisions.

Our work extracts information from existing policies and does not require machine readable language implementation. Additionally, compared to existing work, our work tries to minimize use of patterns or rules because the format of policies often changes over time and rules and patterns often become obsolete over time.

## 2.6 Privacy policy visualization

Visualizing the privacy policies is one of many approaches proposed by several studies aiming to improve privacy policies usability and effectiveness to enhance users' understanding of privacy practices. One of the proposed presentations of privacy policy is the P3P Expandable Grid, an active tool that allows its users to navigate and discover the privacy policy of websites.

It uses websites' P3P specifications to present the information in a holistic view. The P3P Expandable Grid consists of header and body sections. The header displays legends that describe the symbols used in the body section and expandable column headers that describe how information is used and who uses it (Reeder et al., 2008). Figure 1 shows the layout of P3P Expandable Grid.

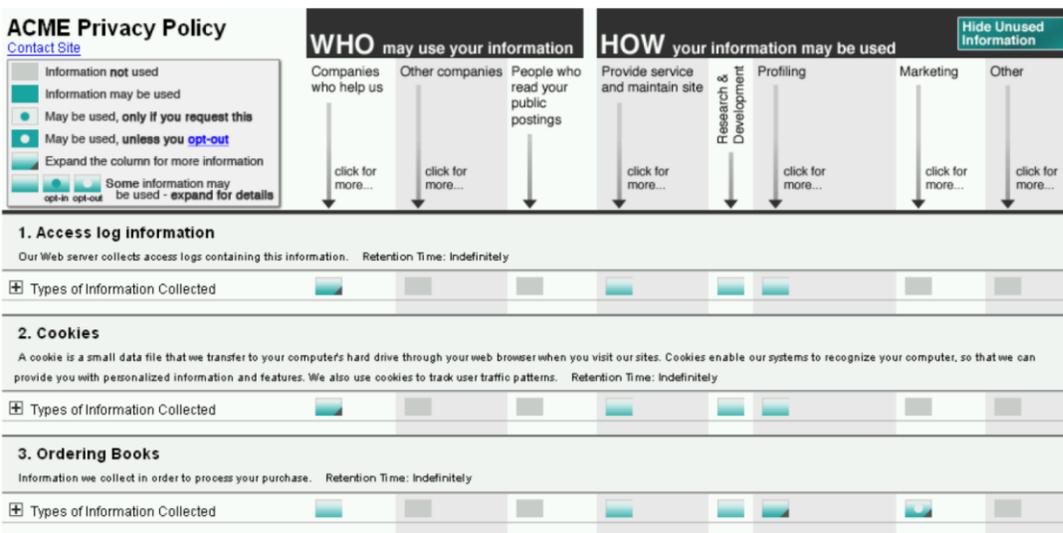


Figure 1: P3P Expandable Grid

Kelly et al. (2009) proposed a “nutrition label” as a standardized presentation for online privacy policies. The standard design is derived from other designs including nutrition, warning, and energy labeling, as well as from the standardized banking privacy notification. The goals of that solution were to improve users’ understanding of privacy policies and to help them find information about their privacy easily. The “nutrition label” shows in a tabular format the type of information use, how this information is used, the purpose of using the information, and with whom it will be shared based on the information stated in the privacy policy. They used

different colors and symbols to present the information. Figure 2 shows the layout of “nutrition label.” Some users found “nutrition label” easy and were satisfied with it. However, one of the limitations is that some other users found the presented symbols confusing and hard to understand. For example, some users were not able to understand what each color meant and what the presented labels meant. Another limitation is that it requires a website to be P3P-enabled to develop the tabular version.

## Acme

information we collect	ways we use your information				information sharing	
	provide service and maintain site	marketing	telemarketing	profiling	other companies	public forums
contact information		opt out	opt out			
cookies						
demographic information		opt out	opt out			
preferences		opt out	opt out			
purchasing information		opt out	opt out			
your activity on this site		opt out	opt out			

**Information not collected or used by this site:** social security number & government ID, financial, health, location.

#### Access to your information

This site gives you access to your contact data and some of its other data identified with you

acme.com

5000 Forbes Avenue  
Pittsburgh, PA 15213 United States  
Phone: 800-555-5555  
help@acme.com

**How to resolve privacy-related disputes with this site**  
Please email our customer service department



we will collect and use your information in this way



we will not collect and use your information in this way



by default, we will collect and use your information in this way unless you tell us not to by opting out



by default, we will not collect and use your information in this way unless you allow us to by opting in

Figure 2: Nutrition Label

Law firm Hunton and Williams introduced a standardized version of privacy policies (Multi-Layered Notices Explained, 2004). It is a short form of a privacy policy that links to the original full text for details. Although it tends to be in a standardized format, only the tabular view and the section headers are standardized. The design of the layered privacy policy can be seen in Figure 3. However, other design details such as the content of the sections and the amount of information to be included are left to be determined by the company. The main weakness of this privacy format is that some companies may use the same language involved in the full natural language policy. This was shown by McDonald et al. (2009), who did not find strong differences between a layered and a full text format when they examined users'

knowledge about some privacy practices.

<b>Acme Privacy Summary</b>	
<b>Scope</b>	This policy discloses what information we gather about you when you visit any of our Web sites (all acme.com and Acme Network sites) or buy product directly from us. For more details, please refer to our <a href="#">full privacy policy</a> .
<b>Personal Information</b>	<p>Acme collects two kinds of information about users:</p> <ol style="list-style-type: none"> <li>1. data that users volunteer by signing up to receive news and product information, entering contests, completing surveys, or buying directly from us</li> <li>2. aggregated tracking data we collect when users interact with us, such as access logs and web cookies</li> </ol> <p>For more information about our information collection practices, please see our <a href="#">full policy</a>.</p>
<b>Uses</b>	<ul style="list-style-type: none"> <li>• We use the personal information you provide voluntarily to send information you've requested and to fulfill orders.</li> <li>• When you sign up online to receive Acme Network newsletters, Acme product and company news, and to participate in talkbacks on our sites you must provide your name, email address, and a password. We never sell or rent your email address or other personally identifiable information you provide us under these circumstances.</li> <li>• When you register for an Acme conference, or sign up for a conference email list, we will send you email announcements and updates about Acme conferences. We send conference brochures to past conference attendees.</li> <li>• When you order books directly from us, or request book catalogs, we add you to our snailmail list, and we'll send you catalogs and other marketing pieces.</li> <li>• When you enter a contest or sweepstakes, we may ask for your name, address, and email address so we can administer the contest and communicate with entrants about the results.</li> </ul>

Figure 3: Layered privacy policy

Most existing work only shows the content of one policy. Our approach also uses a simple table format to compare policies. We also propose a “learn to rank” approach to assign a privacy score that can be used to compare many policies at a time.

## CHAPTER 3: Analysis of User's Privacy Concerns

### 3.1 Introduction

As it becomes easy and inexpensive to store a vast amount of data, information collected by service providers about individuals is growing rapidly. At the same time, concerns about the privacy of individuals are increasing. Although most service providers use privacy policies to explain to users what information they are collecting, who will access it, and for what purpose, existing research shows that users often do not read privacy policies or they find privacy policies difficult to understand. There has been relatively little work done on how users feel about the collecting or sharing of specific information, despite its being vital for organizations to design better privacy policies. In this chapter, we first break down privacy policies of two well-known social network companies, Facebook and Twitter, and one financial institution, Bank of America, into easy-to-understand segments. We then discuss how we conducted a survey using crowd sourcing to analyze users' responses to these policy segments. We asked questions regarding users' awareness, expectations, and privacy concerns for these policy segments. We then investigated the relationship between various factors such as demographic factors and factors related to the privacy policy segments such as data type, data flow, user controls, and users' privacy concerns. We close the chapter with guidelines and suggestions for privacy policy improvements and ways to increase users' awareness of privacy policies.

### 3.2 Breaking down privacy policies

Because privacy policies are usually written in a lengthy and unclear format, we first broke down privacy policies into easy-to-understand policy segments. We created these segments by analyzing the privacy policy text of different service providers represented in a table format to make it easier for us to design our study. The privacy policy segment is a collection of information that describes what information will be collected or shared, the parties involved in the collection or sharing practice, the purpose of that practice, and the control over that practice.

A segment is divided into five fields: data, from, to, purpose, and user control. The “data” field includes the information collected, stored, or shared by the service provider. The “from” field describes the information provider such as other users, service providers, and third parties from whom information is collected. The “to” field includes the party who is receiving the collected or shared information including other users, service providers, and third parties. The “purpose” field describes the reasons for collecting, storing, or sharing stated information. Finally, the “user control” field states whether the user can control the collecting, storing, or sharing of the specified information.

To illustrate clearly the segments and their fields, Table 1 shows how we converted Bank of America’s (BoA) privacy policy into segments with each segment allotted its own line.

The first segment shows that personal information is collected from the customer and then passed to BoA for a purpose such as opening an account or performing transactions,

applying for a loan or using a credit card, and seeking advice about investments. The table very clearly shows that users have no control over the sharing or collecting of these data.

Table 1: Policy segments from Bank of America

	Data	From	To	Purpose	User Control
1	Personal information	Customer	Bank	open an account or perform transactions, apply for a loan or use your credit or debit card, seek advice about your investments	NO
2	Personal information	Bank	Bank	process your transactions, maintain your account(s)	NO
3	Personal information	Bank	Legal request	respond to court orders and legal investigations, or report to credit bureaus	NO
4	Personal information	Bank	Service providers	offer our products and services for customer	NO
5	Personal information	Bank	Financial companies	for joint marketing	NO
6	Information about transactions and experiences	Bank	Affiliates	for everyday business purposes	NO
7	Information about creditworthiness	Bank	Affiliates	for everyday business purposes	YES
8	Credit card accounts	Bank	Nonaffiliates	to market the customer	YES
9	Accounts and services endorsed by another organization	Bank	Nonaffiliates	to market the customer	YES

Using the information in the table we categorized the values of the “data” field of all the privacy policies into four categories: identity information such as name and credit card account, activities information such as user’s interaction with links and messages he sends, log information such as IP address and pages visited, and cookies. We also generalized the values of the “from” and “to” fields into three categories: users, service providers, and third parties. Based on this categorization we created a data flow diagram where each generalized party category forms an element including user, service provider, and third party.

We then represented policy segments using a data flow graph. Each node shows a generalized “from” or “to” field value. Each arrow shows a policy segment where data are passed from a “from” value to a “to” value. The number next to the arrow represents the number of such segments. Figure 4 shows the flow diagram based on the contents of Table 1.

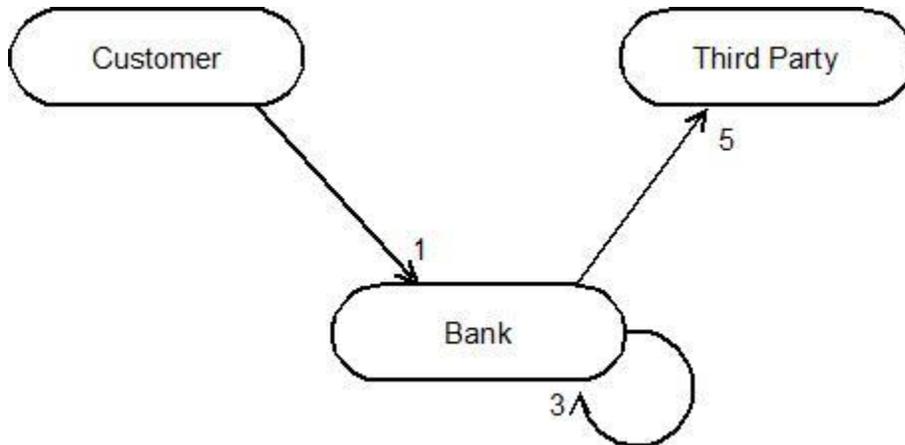


Figure 4: Data flow graph for Bank of America

### 3.3 Research questions

Our study survey served to answer the following questions:

1. Does users' demographic information affect their privacy concerns related to information collection and sharing?
2. Does users' frequency of usage and awareness or understanding of the service provider's privacy policy affect their privacy concerns?
3. Do users prefer additional control over the sharing and collecting of their data?
4. Are users aware of provided controls? Do they use them?
5. Which data flow or data type is most unexpected by users, and with which data flow or data type are users most uncomfortable?
6. Do users have difficulty identifying the reasons service providers collect or share certain types of information?
7. Does clarifying the purpose of collection or sharing information ease users' concerns about the privacy of their data?
8. Is it possible to predict the level of privacy concerns for a user?
9. What other factors affect users' privacy concerns?

### 3.4 Study design

We selected privacy policies from three service providers: Facebook, Twitter, and BoA. We selected Facebook and Twitter based on their rating as the top two social networking websites according to [Ebizmba.com](http://Ebizmba.com). BoA was selected to represent financial institutions because it is rated one of the largest banks in the United States according to [Relbanks.com](http://Relbanks.com).

We designed the surveys to gather information related to all the segment fields. We used Quicksurveys.com to design six surveys, two survey versions for each of the three providers. The two survey versions are identical in terms of the sections and questions except that one version explains the purpose of the data collection to the participants while the other does not. Instead, in the other survey version, participants were asked to identify the reason for which their data were being collected or shared. Comparing the results of these three pairs of surveys allowed us to test the hypothesis that informing users of the purpose of data collection and sharing practices eases their privacy concerns.

The surveys were divided into the following sections:

- The first section gathers the participant's demographic information.
- The second section verifies that the participant has an account with the specific service provider and asks how frequently the participant accesses and uses the account.
- The third section assesses the participants' awareness or understanding of providers' privacy policy, to which they agreed before creating their accounts.
- The fourth section has two different sets of questions in each version of the survey. The first set of questions, called the "purpose condition," asks participants to indicate their

level of comfort knowing that the service provider is collecting or sharing their information for a specified reason. For instance, Twitter tracks users' interactions with links to help them improve their service and deliver relevant advertisements.

The second set of questions, called the "expectation condition," asks the participants to identify the reason a service provider is collecting or sharing specific information and then asks the participants' comfort level knowing this data will be collected or shared without explaining the purpose of the collection or sharing. A 5-point Likert scale ranging from very comfortable (+2) to very uncomfortable (-2) was used to measure the participants' comfort levels. Figure 5 shows some sample questions of the two conditions.

- The fifth section asks participants if they want to have control over their data collection and sharing practices.
- The sixth section asks participants if they are aware of the controls the service provider makes available to the user related to data collection and sharing practices, and, if so, whether the participants use the provided controls.

<p><b>Q</b> 3. Could you think of any reason that Twitter tracks your interaction with links (clicking on links)?</p> <p><b>A</b> I cannot think of any reason</p> <p>To help them improve their services and provide more relevant advertisements</p> <p>To help others find your account</p> <p>Sharing of the information</p> <p>Other, please specify</p> <input type="text"/>	<div style="border: 1px solid black; padding: 2px;">Expectation Condition</div>
<p><b>Q</b> 4. Are you comfortable when Twitter tracks your interaction with links (clicking on links)?</p> <p><b>A</b> Very uncomfortable</p> <p>Somewhat uncomfortable</p> <p>Neutral</p> <p>Somewhat comfortable</p> <p>Very comfortable</p>	
<p><b>Q</b> 3. Are you comfortable when Twitter tracks your interaction with links (clicking on links) to help them improve their service and provide more relevant advertisements?</p> <p><b>A</b> Very uncomfortable</p> <p>Somewhat uncomfortable</p> <p>Neutral</p> <p>Somewhat comfortable</p> <p>Very comfortable</p>	<div style="border: 1px solid black; padding: 2px;">Purpose Condition</div>

Figure 5: Sample questions

We published the surveys through the crowd sourcing tool Amazon’s Mechanical Turk (AMT) because of crowd sourcing’s low cost and its ability to attract many participants from diverse backgrounds. We designed a human intelligence task (HIT) to link to each survey. Each HIT contained a short description of the survey, a link to the related survey, and a code to be entered after completion of the survey for verification and approval.

We made a between-subject study design in which different participants participated in each condition. We posted 11 different HITs, two for each service provider and each condition. For the Twitter expectation condition, however, only one HIT was posted because it was our first published survey and we chose 50 participants for that HIT, which made it hard to manage and approve; thus, we decided to design the other HITs to accept only 25 participants. On average, each HIT took 12 days to be completed. Participants spent about 3 minutes and 42 seconds per HIT and were paid at the rate of \$1 per HIT.

We collected a total of 323 responses, 93% of whom were American participants. Tables 2 and 3 show the gender and age distribution of the participants. Twenty of the submitted responses were discarded due to the participants' rejection of the consent form or because they did not have an account with one of the study's service providers.

*Table 2: Age distribution of the participants*

<b>Age range</b>	<b>18-25</b>	<b>26-30</b>	<b>31-35</b>	<b>36-40</b>	<b>40+</b>	<b>Total</b>
<b>Number of participants</b>	61	74	51	51	86	323
<b>Percentage</b>	19%	23%	16%	16%	27%	100%

*Table 3: Gender distribution of the participants*

<b>Gender</b>	<b>Females</b>	<b>Males</b>	<b>Total</b>
<b>Number of participants</b>	183	140	323
<b>Percentage</b>	57%	43%	100%

### 3.5 Results

When analyzing the data, we focused on two main categories, both with specific subcategories. The first category, data type, has four subcategories: personal, activity, log, and cookies information. The second category, data flow, has three sub-categories: user-X, X-Thirdparty, and Thirdparty-X in which X is a service provider.

#### 1. Does users' demographic information affect their concerns?

The average comfort level of our data shows that the older the participants were, the more concerned they were about the privacy of their information, with averages of (-0.6) and (-1.01) for age groups 18–25 and 40+, respectively. The results are statistically significant with a 0.006 p-value in the t-test. In addition, we calculated the average privacy concerns based on participant gender. The average concern of female participants was (-0.92) and of males was (-0.67), indicating that females are more concerned about the privacy of their information than males. These results are also statistically significant because the p-value is 0.008. Figures 6 and 7 show the average comfort levels of participants based on their age and gender.

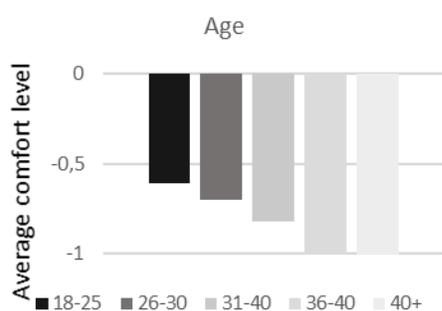


Figure 6: Average comfort level based on age

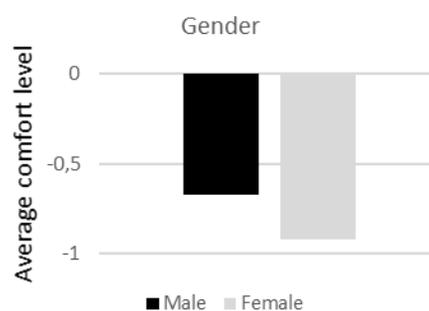


Figure 7: Average comfort based on gender

2. Do users' frequency of usage and awareness of the service privacy policy affect their privacy concerns?

We calculated the average of participants' comfort levels regarding the frequency with which they use certain services. The averages fall between (0.75), for multiple times a day usage, and (0.93) for weekly usage. The results are shown in Figure 8. Based on t-test results, in which we can conclude that the results are not statistically significant, the p-value is 0.43.

Overall, the participants' awareness of the privacy policies of services they use is low; 65% of the participants were only partially aware of the privacy policies, and 23% of participants were not aware of privacy policies. The results show that participants who did not know of the existence of the privacy policies were more concerned about their privacy than the participants who were aware of the privacy policies. The average comfort level of participants who did not know service providers have privacy policies is (-1.8), and the average comfort level of participants who were fully aware of the privacy policies is (-0.53). The results are shown in Figure 9. The difference is statistically significant with a p-value of 0.0005, which indicates that awareness of the content privacy policies decreases participants' privacy concerns.

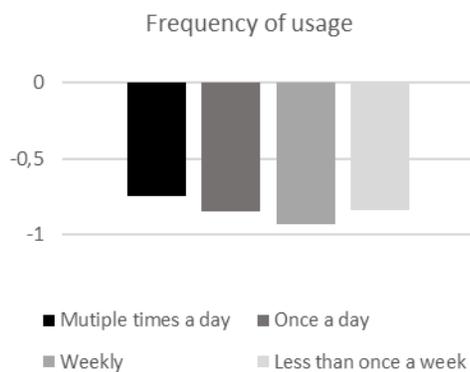


Figure 8: Average comfort based on frequency of usage

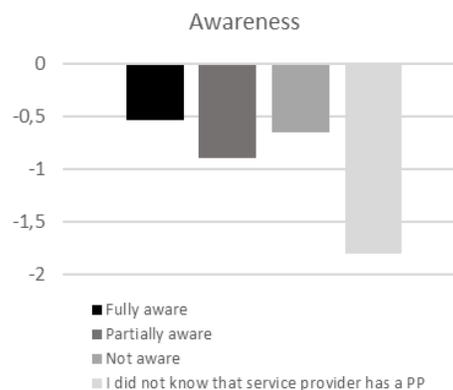


Figure 9: Average comfort level based on awareness of the privacy policy

### 3. Do users prefer more controls?

The clear majority of participants prefer to control their own data collection and sharing practices. This is clearly seen by the fact that 92% of the participants thought that service providers should allow them to control the sharing or usage of their own information. Table 4 shows more details about users' control preferences.

Table 4: Percentage of control preferences

Do you want more control over your privacy?			
Service provider	Yes	No	Does not matter
Facebook	90%	3%	7%
Twitter	92%	1%	7%
Bank of America	93%	6%	1%
<b>Average</b>	<b>92%</b>	<b>4%</b>	<b>4%</b>

#### 4. Are users aware of the provided controls? Do they use them?

About 62% of participants were not aware of the controls offered by the service providers. However, 63% of those who know that they have certain controls use them. Tables 5-7 show the detailed results for three companies.

*Table 5: Percentage of participants' awareness of and usage of Facebook existing controls*

Are you aware of controls provided by Facebook?	
<b>Yes</b>	<b>No</b>
40%	60%
Do you use them?	
<b>Yes</b>	<b>No</b>
69%	31%

*Table 6: Percentage of participants' awareness of and usage of Twitter existing controls*

Are you aware of controls provided by Twitter?	
<b>Yes</b>	<b>No</b>
52%	48%
Do you use them?	
<b>Yes</b>	<b>No</b>
57%	43%

*Table 7: Percentage of participants' awareness of and usage of Bank of America existing controls*

Are you aware of controls provided by Bank of America?	
<b>Yes</b>	<b>No</b>
21%	79%
Do you use it?	
<b>Yes</b>	<b>No</b>
64%	36%

5. Which data flow or data type is the most unexpected by users, and which data flow or data type are users most uncomfortable with?

Using the “exception condition” version of the survey, we analyzed the results based on the data type categories and data flow categories to investigate which category is least expected by the participants.

First, for each service provider, we calculated the percentage of the participants who expected the collection or the sharing of their information through each flow. We also averaged the self-reported comfort ratings, ranging from “very comfortable” at +2.0 to “very uncomfortable” at -2.0, with “neutral” at 0.

We observed a strong correlation ( $r=0.76$ ) between the percentage of expectations and the average comfort ratings. For example, 28% of participants expected Twitter would share their information with a third party, and overall participants felt uncomfortable about this information sharing (-1.26). However, 94% of participants expected a third party would share their information with Facebook. Overall, the average comfort level of (-0.54) shows that participants felt somewhat comfortable about this information sharing. We can conclude that when participants were not expecting a collection or sharing of their data through particular flow, they were less comfortable with this practice. The summary of our findings is shown in Table 8.

Table 8: Percentage of expectation and average comfort level based on type of flow

Type of flow	Service provider	% Expectation	Average comfort
user-X	Facebook	-	-
	Twitter	43%	-0.62
	BoA	-	-
X-Third Party	Facebook	42%	-0.89
	Twitter	28%	-1.26
	BoA	42%	-1.00
Third Party-X	Facebook	94%	-0.54
	Twitter	27%	-1.00
	BoA	-	-

Next, we analyzed the results for each service provider and data type category. We calculated the percentage of the participants who expected the collection or the sharing of a data type. We also averaged the participants' self-reported comfort ratings.

We observed a correlation of ( $r=0.5$ ) between the percentage of expectations and the average comfort ratings; 95% of Facebook users expected that Facebook collects or shares information about their activity information, and overall results showed that Facebook users felt somewhat comfortable about Facebook's monitoring their activities (-0.54); 47% of BoA users expected that BoA collects or shares information about their activity information, and overall results showed that BoA users felt uncomfortable about BoA's monitoring their activities (-1.2). The findings indicate that the expectation of collection or sharing of a specific data type was somewhat linked to participants' subjective feelings. The more they expect the collection of a specific information the more comfortable they feel, and the less they expect the collection of a specific information the less comfortable they feel. The summary of the findings is shown in Table 9.

Table 9: Percentage of expectation and average comfort level based on type of data

Type of data	Organization	% Expectation	Average comfort
Personal	Facebook	52%	-0.81
	Twitter	34%	-1.05
	BoA	37%	-0.8
Activity	Facebook	95%	-0.54
	Twitter	61%	-0.41
	BoA	47%	-1.2
Log	Facebook	-	-
	Twitter	20%	-0.56
	BoA	-	-
Cookies	Facebook	10%	-1.18
	Twitter	-	-
	BoA	-	-

#### 6. Do users have trouble identifying the reason service providers collect and share information?

We found that participants had trouble identifying why service providers collect or share their information, with an average of 43% guessing correctly.

We divided the purposes of privacy policy segments into three categories: (a) for major functionality, (b) for sharing and tagging, and (c) for target advertising or market analysis. Some privacy purposes fell into more than one category. For example, the purpose of Facebook's collecting users' activities falls into both the major functionality and advertising categories.

We compared the provided purposes for information collection and sharing of the participants against the actual purposes as stated in privacy policies for different types of data categories. The results as shown in Table 10 indicate that most the participants could not correctly identify why the service provider collects or shares a specific type of data. While more participants could identify why their activity information was collected or shared compared to

their personal, log, and cookies information, overall, less than 50% selected the correct reason for these service providers' practices.

Table 10: Type of data

<b>Type of data</b>	<b>Information used for:</b> <b>[1] Major functionality</b> <b>[2] Tagging or sharing</b> <b>[3] advertising or market</b>	<b>% of correct choice</b>	<b>% of do not know</b>
Personal	[1]	47%	18.9%
	[2]	37%	26.4%
	[3]	37%	10.2%
Activity	[1] + [3]	78%	2.9%
	[3]	47%	20.4%
Log	[1]	20%	13%
Cookies	[1]	10%	10%

In this table, the first column shows type of accessed data. The second column shows the ground truth of why data is shared. The third column shows the percentage of participants stated the purpose correctly. The last column shows the percentages of participants who chose that they cannot think of a reason.

We followed the same analysis we performed on type of data for different categories of data flow. The results show that only 48% of the participants on average selected the correct reason for information collection and sharing practices. We noticed that for third party-X type of flow an average of 59% participants chose the correct answers. This average is better than the other two categories of data flows in which only 40% of answers were correct for X-Third Party and only 48% of answers were correct for user-X. Table 11 shows all the results.

Table 11: Type of flow

Type of flow	Information used for [1] Major functionality [2] Tagging or sharing [3] advertising or market analysis	% of correct choice	% of do not know
user-X	[1]	61%	13.0%
	[1] + [3]	34%	1.9%
X-ThirdParty	[1]	41%	32%
	[2]	37%	26.4%
	[3]	42%	15.3%
ThirdParty-X	[1]	94%	4%
	[1] + [3]	24%	25.9%

In this table, the first column shows the type of flow. The second column shows the ground truth of why the data are shared in this type of flow. The third column shows the percentage of participants stated the purpose correctly. The last column shows the percentages of participants who chose that they cannot think of a reason.

7. Does clarifying the purpose of the information collection or sharing ease users' concerns about the privacy of their data?

In this section, we study whether clarifying the purpose of the information collection or sharing eases users' privacy concerns. We compared the average comfort ratings from surveys using an expectation condition in which the purposes of information collection or sharing are not revealed to participants to the average comfort ratings from surveys using purpose condition in which the purpose is revealed. The results were analyzed based on data type and data flow categories.

The results show that the differences between the comfort ratings with and without stating the purpose of information collection and sharing were not statistically significant. For example, regarding activity type data, the t (298) and p-value are 1.97 and 0.102, respectively. This and the other results can be found in Table 12.

*Table 12: Comparison of comfort ratings between the expectation condition (2<sup>nd</sup> column) and the purpose condition (3<sup>rd</sup> column) based on data type.*

Data type	Comfort rating w/purpose	Comfort rating <b>w/o</b> purpose	df	T	P
Activity	-0.51	-0.71	298	1.97	0.102
Personal	-1.02	-0.88	299	1.97	0.213
Log	-0.48	-0.56	102	1.98	0.720
Cookies	-1.08	-1.18	99	1.98	0.573

Table 13 shows the results for the three different types of data flow. The difference between the comfort ratings was not statistically significant, as shown by the example of user-X data flow, in which the t (102) and p-value are 1.98 and 0.402, respectively.

*Table 13: Comparison of comfort ratings between the expectation condition (2<sup>nd</sup> column) and the purpose condition (3<sup>rd</sup> column) based on type of flow*

Type of flow	Comfort rating w/purpose	Comfort rating <b>w/o</b> purpose	df	T	P
User-X	-0.47	-0.62	102	1.98	0.402
X-ThirdParty	-1.11	-1.06	294	1.97	0.632
ThirdParty-X	-0.68	-0.78	202	1.97	0.509

#### 8. Is it possible to predict the level of privacy concerns for a specific user?

In this section, we proposed a model using logistic regression to predict the level of concern for a user with specific demographic information, specific frequency of use, and specific level of awareness. We used R in our study as an analytical tool. We had one dependent

variable, comfort level, and seven independent variables: service, dataflow, datatype, gender, age, frequency of use, and awareness level. These variables are either categorical or ordinal. Table 14 shows each variable type. We started first by analyzing the data of all 605 instances for the three service providers. The results showed a value of 0.19 for the  $R^2$ , indicating a weak model. To improve our model and remove the data variant in concerns, we divided our data into two clusters based on the comfort level variable. We ran k-means clustering with  $k=2$  and got 383 instances that belong to Cluster1 in which participants were more concerned (i.e., lower comfort level) and 222 instances that belong to Cluster0 with participants who were less concerned. We then applied logistic regression on each cluster. The results show an  $R^2$  value of 0.28 for Cluster1 and 0.26 for Cluster0. We also applied different combinations of the data to obtain better results with higher  $R^2$ . Table 15 shows the  $R^2$  values of each combination.

Table 14: Variables type

Variable	Type
Organization	Categorical
data_flow	Categorical
data_type	Categorical
Gender	Categorical
Age	Categorical
freq_usage	Categorical
Awareness	Categorical
Comfort level (dependent)	Ordinal (-2, -1, 0, 1, 2)

Table 15:  $R^2$  values for the results of logistic regression modelling

	All participants	Cluster 0	Cluster 1
All three companies' data	0.19	0.288	0.268
Data for Facebook only	0.179	0.33	0.272
Data for Twitter only	0.313	0.349	0.478
Data for Facebook and Twitter	0.134	0.485	0.206
Data for BOA only	0.457	0.069	0.438

We also employed linear regression and ordinal regression, but neither method provided better  $R^2$  results. Table 16 shows the  $R^2$  values of each combination using the linear regression method. In most cases, the results were worse than applying logistic regression in terms of  $R^2$  values. According to the  $R^2$  values of the regression models, we found that it is difficult to predict a participant's privacy concerns given specific values by applying regression models.

In the next section, we show which independent variables correlate to users' privacy concerns.

*Table 16:  $R^2$  values for the results of linear regression modelling*

	All participants	Cluster 0	Cluster 1
All three-company's data	0.163	0.272	0.228
Data for Facebook only	0.124	0.448	0.175
Data for Twitter only	0.268	0.326	0.422
Data for Facebook and Twitter	0.155	0.307	0.233
Data for BOA only	0.379	0.064	0.329

### 9. What other factors affect users' privacy concerns?

To study other factors that affect users' privacy concerns, we used an R statistical tool to build a logistic regression model. Table 17 shows independent attributes along with p values and coefficients that have significant correlation with comfort levels. Note that all independent variables are categorical, so R changes them into factors so that each value of a categorical variable is a binary variable. For example, each category of data flow will become a binary variable.

The results show that some factors related to data flow, data type, type of organization, and lack of awareness of privacy policy affect users' privacy concerns. For example, Facebook is positively correlated with comfort levels, which may be because people are willing to share their information on Facebook. Users are generally fine when their data are shared with the service provider. Users are concerned when cookies or personal data is being collected and shared. Users who are partially aware or unaware of the privacy policy also have greater privacy concerns.

The rest of the attributes' correlation results support the findings of previous research questions 1 and 2. Table 17 shows these correlations with p-values <0.05. For example, older people and females are more concerned with privacy. Also, a lack of awareness of privacy policies increases privacy concerns.

*Table 17: Independent attributes that have significant correlation with comfort levels*

<b>Attribute</b>	<b>P-Value</b>	<b>Coefficient</b>
organization=social page (Facebook)	<0.0001	1.09
data flow=user-X	0.007	0.78
data type=cookies	0.002	-1.1
data type=personal	0.003	-0.67
gender=male	0.0005	0.58
age=36-40	0.001	-0.85
age=40+	0.001	-0.76
awareness=did-not-know-X-has-PP	0.004	-1.4
awareness=Partially-aware	0.002	-0.84

### 3.6 Some guidelines

Based on our findings, we came up with a list of guidelines to help organizations improve their privacy policies.

1. Improve privacy policy readability and describe privacy practices more clearly, especially regarding the use of cookies, personal information collection, and the sharing of users' information with third parties. This is suggested based on the finding that a large portion of users are partially unaware of some policy segment fields and thus they tend to have higher levels of privacy concerns.
2. Make it more explicit to users that they have controls on certain information usage or sharing. This is based on the finding that most users are unaware of some provided controls; however, the users who know and understand these controls tend to use them.
3. Develop a visualization model for the privacy policies using the flow chart in Figure 4.
4. Another possible suggestion is to use tools that help people understand existing privacy policies. These tools use natural language processing techniques to parse automatically or semi-automatically existing privacy policies and present them in easy to understand formats such the table format we used in the chapter.
5. Develop a comparison model of privacy policies to help users better determine which policy is preferable to them based on the information it collects, the party with whom it shares information, and the controls it provides.

### 3.7 Discussion and Conclusion

The purpose of this study is to analyze users' responses to privacy practices presented in the privacy policy of online social networks and financial institutions. We started by breaking down privacy policies into easy-to-understand segments and then used crowd sourcing to collect data.

The results of this study show that demographic factors do affect users' levels of concern related to the privacy of their information. It shows older users have greater concerns about their privacy. This finding supports Hazari and Brown's (2013) argument that privacy concern increases with age. Our study results indicate that females are more concerned about their information privacy, which supports the position of Nowak and Phelps (1992) but contradicts Hazari and Brown (2013), who stated that gender difference is not a significant factor in determining users' levels of privacy concern.

In addition, our study shows that an average of 92% of the participants preferred that companies allow users to control their data collection or sharing. The results also show that most users who know that they have control over some of their information will use the provided controls. This finding agrees with other findings (Consumer Action Organization, 2013; Young & Quan-Haase, 2009) that have stated that privacy control is a critical issue for users. However, surprising results of our study show that, when controls are provided, not all users are aware of them. We found an average of only 38% of participants were aware of provided controls. So, we believe service providers need to make information on such controls more prominent and explicit.

Our results suggest that when users do not expect their data to be collected or shared, especially when involving third parties, log data, or cookies, they tend to have more privacy concerns. We suggest that privacy policies be written to make such practices more obvious to users. Nonetheless, while our original finding is consistent with Lin et al.'s (2012) results, our results show that providing the reason for the collection and use of particular information collected or used had no significant influence on users' privacy concerns, while their results show purpose has a significant impact on users' concerns. One possible reason could be that in our settings of online social networks and financial services, participants already know that their information is being collected or shared, and they are expecting that this is done for specific reasons. In their settings of smart phone apps, users are unaware of what data are shared or collected in the first place.

We also built a regression model to predict the level of privacy concerns for a specific user for a specific privacy policy segment field. This proved to be a weak model. A reason could be that our model did not capture other factors affecting users' concerns such as whether the user is concerned with privacy in general. This model indicates that, in addition to demographic factors, the type of organization also affects users' privacy concerns. For example, Facebook users are less concerned with privacy than those of the other two organizations. A possible reason is that Facebook users mainly use the service to post information about daily activities, thoughts, pictures, and interests and to share that information with others, so they felt more comfortable regarding the privacy of their information. However, this is not the case with BoA, as the main goal of using the service is not sharing information. The situation is similar for

Twitter, whose users are more likely to use the service for posting news and comments with 140-character messages.

Other factors that affect privacy concerns include data type, in which users are more concerned with sharing cookies and personal data and data flow in which users' concerns are higher when data are shared with third parties. Another remarkable result is that the lack of awareness of some policy practices is strongly correlated with increased privacy concerns.

## CHAPTER 4: Information Extraction

### 4.1 Introduction

Although privacy policies are used by companies to communicate their data collection and sharing practices to users, many studies show that the policies are difficult to comprehend or otherwise hard to read. The results of our study on users' privacy concerns showed that some users are not aware of and therefore concerned about certain privacy practices such as the collection and sharing practices, their preferences for more control over these practices, and their unawareness of the applicable options.

Instead of requiring the users to read entire policies written in natural language to understand privacy practices, we believe presenting a summary of the practices would effectively improve their awareness about privacy practices. This will require us to analyze the policies written in natural language and extract data related to privacy practices.

Our goal in this chapter is to propose extraction techniques that automatically extract data related to collection, sharing, and control practices from the privacy policies. We will study and analyze the privacy policies of randomly chosen service providers from different domains. Our focus will be on how collection, sharing, and control practices are described within the policy text and on common text characteristics related to these practices.

Constante et al. (2012) proposed an information extraction solution to show what personal information is collected by automatically analyzing privacy policy text. Although the solution could extract more than 80% of the information, it presented some limitations. The

extraction is based on pattern definition and rule creation, which require human interaction. The system's extraction accuracy depends on a part of speech (POS) tagger. The following sentence from the study provides an example of the limitations of the extraction system in which relevant information was not collected: "LinkedIn requests other information from you during the registration process, (e.g. gender, location, etc.)." The word "requests" can be a singular verb or a plural noun in the English language and as such "requests" was not recognized as a verb by the POS tagger. This means that the sentence did not fit the predefined patterns and the words "gender" and "location" were not be identified as personal information.

In our solution, we will avoid the reliance on the speech tagger, definition of patterns, and rule creation, especially when the searchable information can be specified by a list of terms. Rather we will analyze the policies and use the analysis results to create a list of terms related the practices we are studying and then search for these terms within the policies. For the hard-to-list practices we plan to extract the information based on common text characteristics.

## 4.2 Information extraction approach

### 4.2.1 Data collection

We formed our dataset from 24 web privacy policies from different domains: health care, financial, social networking, shopping, and other. Several domains were included to ensure that the observation discovered during privacy analysis are shared between a wide range of providers and can be generalized. The domains and providers are shown in Table 18.

Table 18: List of domains and service providers

Domain		Service Provider	
1	Health care	1	CIGNA
		2	Mayo Clinic
		3	Quest Diagnostics
		4	Humana
		5	Massachusetts General Hospital (Mass)
2	Financial	6	Bank of America
		7	Citibank
		8	Visa
		9	Chase
		10	Western Union
3	Social networking	11	Facebook
		12	Twitter
		13	Instagram
		14	LinkedIn
		15	Tumblr
4	Shopping	16	Bloomingdale's
		17	Neiman Marcus
		18	Net-A-Porter
		19	Barney NY
		20	Shopbop
5	Others	21	Kemper
		22	Hotels
		23	Priceline
		24	Social Security Administration (SSA)

#### 4.2.2 Preprocessing

The text of the privacy policies in our dataset was cleaned by removing the extra spaces and bullet points and moving the bulleted information into the appropriate paragraphs. Then it was saved in a text file format.

The dataset is divided into two groups. The first is a training set that includes three privacy policies from the four specific domains for a total of 12 policies. The second group is a

testing set with 12 privacy policies, two from each specific domain and four policies from other random domains. We split the dataset into these two sets to avoid bias when measuring the extraction accuracy. While the testing dataset will not be used in either the analysis or the validation process, the training dataset will be used to analyze the data and to develop the extraction algorithm. In other words, we analyzed the training dataset to find common words used in describing different practices, used the analysis results to develop the extraction algorithm, and then evaluated the algorithm using the testing dataset. Further, word stemming, partial match, and negative handling were taken into consideration during this step.

#### 4.2.3 Personal information

Our definition of “personal information” is any information about an individual that (a) can be used to identify an individual such as name, biometric records, or social security number; (b) can be used to trace the online activities of an individual such as an IP address or audit log; or (c) is linked or linkable to other information about an individual including financial information, place of birth, and race.

Our definition of “personal information” is based on the previously determined and provided definitions in several privacy-related laws and regulations. The following are examples of the laws and regulations involved in our definition:

1. Graham-Leach-Bliley Act of 1999 (GLBA)—governs privacy and use of financial information.
2. Health Insurance Portability and Accountability Act of 1996 (HIPAA)—governs privacy and use of health care information.

3. Family Educational and Privacy Rights Act of 1974 (FERPA)—governs privacy and use of student information.

4. North Carolina Identity Theft Protect Act of 2005—governs privacy and use of personal and financial information.

#### *4.2.3.1 Personal information classification-based extraction*

We used classification methods to categorize the privacy policies' contents as personal and non-personal information.

To identify the personal information, we extracted noun phrases and then classified them as either personal or non-personal information. The noun extraction algorithm uses the text file (f) created during the personal information preprocessing step as an input and generates a list of noun phrases (n) as an output. The output list is then used in a classification process.

##### *4.2.3.1.1 Noun extraction*

For the first step in the noun extraction process, the text in file (f) was split into sentences (s). Words in these sentences were then tagged by their parts of speech using a Stanford POS tagger. Next, for each sentence, the words tagged with one of predefined noun phrases tags were added to a list of noun phrases (n). Whenever a word with a tag that does not match the predefined noun phrases tags was reached, it was skipped and the algorithm started to look for a new noun phrases. Finally, the process terminated when last word in the text was reached.

<b>Algorithm 1:</b> Extraction of Noun Phrases
<b>Input:</b> Information collection text file (f)
<b>Output:</b> List of noun phrases (n)
<b>Begin:</b> 1 split paragraph into sentences (s) 2 tag the sentences 4 <b>for</b> each tagged sentence 5 <b>for</b> each word 6         if the tag of the word match noun phrases tags (NN, DT, IN, CC, and JJ) add word to the current noun phrase 7 <b>else</b> skip the word add current noun phrases to the list (n) start new phrase <b>end if</b> <b>end for</b> <b>end for</b> 8 <b>return</b> list (n)
<b>End Begin</b>

#### 4.2.3.1.2 Classification pre-processing

To classify the produced noun phrases into personal and non-personal information we started by transferring the list (n) into a Term-Document Matrix. Each noun phrase represents a document and the words included in all the phrases represent the terms. We used the text-to-matrix generator (TMG) tool to perform this step. A filter to exclude words less than three letters was applied. The output of the text transformation produced a matrix of 836 phrases by 555 words. The data were then labeled manually as either personal information (p) or non-personal information (np).

#### 4.2.3.2 *Personal information rule-based extraction*

##### 4.2.3.2.1 Analysis Process

We analyzed the training dataset to find if there are specific structures or forms of how personal information collection practices is stated in the privacy policies. We found that there are five main components of a sentence stating what the collected personal information is (Costante, Hartog, & Pekovie, 2013). Data collector represents who collects the information, for example the company name. Collector action represents a verb that describes the action that a collector takes such as request, collect, and store. Data provider represents the data owner, usually the user, who's data is collected. Provider action represents the action that a user does when providing his information such as provide, choose, and determine. Personal data represents users' collected information. We created a table to come up with set of rules with columns representing the five components and rows that represent the position of that component in the different sentences as stated in the privacy policies. Table 19 shows an example of some of the created rules based on the following sentences.

1. <sup>1</sup> **We collect** <sup>2</sup> Personal Information such as: **Your name; Contact information; Visa card number.** <sup>3</sup>
2. <sup>1</sup> **We collect** <sup>2</sup> personal information that **you** <sup>3</sup> voluntarily **provide** <sup>4</sup> through our Services, including: **Name, address, and birthdate;** <sup>5</sup> **Other contact information such as email address and/or text address; Financial and health information; Credit or debit card number; and Social security or similar national ID number; Geolocation information; and Social media account IDs.**
3. Your correct <sup>1</sup> **date of birth**, which <sup>2</sup> **we** may <sup>3</sup> **require** that <sup>4</sup> **you** <sup>5</sup> **submit** accurately to ensure that you do not access the Site if you are not of a legal age to purchase items offered for sale on the Site;

Table: 19 An example of rules table

#	Data collector	Collector action	Data Provider	Provider action	Personal data
1	1	2			3
2	1	2	3	4	5
3	2	3	4	5	1

Based on the created rules we could detect the personal information and then identified it.

#### 4.2.3.3 Personal information matching-based extraction

##### 4.2.3.3.1 Analysis Process

We analyzed the training dataset to find where and how the information about personal information collection practices is stated in the policies. We observed the following:

1. A separate section is used to describe what personal information the service providers collect.
2. Most of the text that describes data collection uses a set of common action words such as “collect,” “provide,” and “obtain.” For instance, Visa uses the action word

“collect” when describing information, it collects from users such as in the following statement: “We collect Personal Information (information which can be used to identify an individual) such as: Your name; Contact information (such as e-mail, phone, and mailing address); Visa card number.” Table 20 shows additional action words that Visa and other providers use in their information collection sections. We will examine if the appearance of these action words in the privacy policy text helps to identify when the keyword is being used as personal information and therefore constitutes a relevant match. We will also examine whether irrelevant matches occur when these action words are not presented.

3. All the data described as personal information in the policies’ information collection section fits our previous definition of “personal information.”

Table 20: List of action words

	access	ask	choose	collect	determine	gather	hold	mean	obtain	provide	receive	record	register	reported	store	submit	tell	use
Quest Diagnostics				x						x								
CIGNA				x						x								
Mayo clinic																		
Visa				x														
CitiBank				x		x				x								
Bank of America				x				x										
Neiman Marcus		x		x							x				x			x
Net-A-Porter	x			x			x		x	x						x		
Bloomingdales				x		x				x								
Instagram	x		x	x						x		x	x	x	x			x
Facebook				x														
Twitter		x	x	x	x					x	x				x		x	x

#### 4.2.3.3.2 Extraction preprocessing

Based on our observation that personal information is written in a separate section of a privacy policy, we extracted information only from this section instead of the entire privacy policy. We saved the data collection section of each service provider’s policy in a text file

format. We also created a list of terms related to personal information that was used to match to the same terms in each service provider's policy, thereby determining what personal information is included in the policies.

#### 4.2.3.3.3 Personal information matching-based extraction algorithm

To find the personal information collected by a certain privacy policy we developed a matching-based extraction algorithm. The algorithm uses the text file (f) of the privacy policy's data collection section, created during personal information preprocessing, and a list of personal information terms (l) as the inputs and generates a list of personal information terms (c) as the output. These generated terms match the terms in the file (f). Each time a term from the input list (l) is matched to a paragraph (p) it is added to the output list (c). A match is determined in three steps. First, a paragraph (p) must contain a term from the input list (l). The same paragraph must also contain an action word. Finally, it must be determined that the term is not an exception.

Exception terms include "age," "address," "SIM," and "name" because they can appear in the middle of other irrelevant words leading to the extraction of non-personal information. For example, "age" can be matched to a paragraph that contains the word "usage" or "SIM" can be matched to a paragraph that contains the word "similar." A conditional statement is added to the algorithm to handle these exceptions by checking if the term is a word or part of another word.

<b>Algorithm 2:</b> Personal Information Matching-Based Extraction
<b>Input:</b> information collection text file (f) list of personal information (l)
<b>Output:</b> list of personal information covered in providers' privacy policies (c)
<b>Begin:</b> 1 <b>for</b> each keyword in the list (l) $i \rightarrow 1$ to N 2 <b>for</b> each paragraph (p) in the text $j \rightarrow 1$ to M 3 <b>if</b> paragraph contains keyword 4 <b>if</b> keyword not an exception & p contains action word add the keyword in personal information collection list (c) 5 <b>else</b> skip <b>end if</b> <b>end if</b> <b>end for</b> <b>end for</b> 6. <b>return</b> list of personal information collection (c)
<b>End Begin</b>

#### 4.2.4 Controls extraction

We define “controls” as any information or practice users can control by stopping the correction or sharing or by opting in or out of that practice.

##### 4.2.4.1 Analysis process

We analyzed the training dataset to find how users' choices and options are described in the policies. We found the following observations:

1. Unlike the collection practices, sentences related to users' control options are not included in a separate section. Rather, they are described throughout the policy text.
2. Text that describes options and choices contain “control keywords,” a set of keywords used in privacy policies that refer to users' choices or options, such as: “control,” “choice,” “options,” “opting,” and “unsubscribe.” For instance, Citibank

uses “control” when explaining that their users can control the sharing of cookies as in the following statement: “You can *control* whether to accept cookies or not.”

Table 21 shows additional keywords that Citibank and other providers use in describing users’ options. The appearance of these keywords will be examined in the evaluation process to determine if their presence helps in reducing the retrieval of unrelated data.

3. Providers allow their users to control the collection of sharing their personal information or to opt out from practices related to the following terms: “address book,” “advertisement,” and “offers.” For example, Neiman Marcus allows their users to stop receiving third party advertisements by stating the following: “To choose to opt-out of this service we have with our third-party advertising partner, click here.” Table 22 shows the specific information and practices service providers allow their users to control.

Table 21: List of control keywords

	choices	choose	control	opt in	opt out	opting	options	unsubscribe
Quest Diagnostics		x	x					
CIGNA		x			x			
Mayo		x			x			
Visa	x	x			x			
CitiBank			x		x		x	
BoA	x	x			x	x		
Bloomingdale's			x		x			x
Neiman Marcus	x	x	x	x	x		x	
Net-A-Porter		x			x	x		
Instagram	x	x	x		x			x
Facebook		x	x					
Twitter			x				x	x

Table 22: Control related terms

	address book	advertisement	communication	contacts list	cookies	flash	location	mail	market	newsletter	offers	promotion	research purpose	SMS
Quest Diagnostics								X						
CIGNA		X	X		X			X	X	X				
Mayo		X	X				X			X			X	
Visa		X							X	X		X		X
CitiBank		X			X	X		X	X		X			
BoA		X			X	X			X	X	X			
Bloomingdale's		X			X	X		X	X			X		
Neiman Marcus		X	X		X		X	X	X		X	X		X
Net-A-Porter		X			X			X	X	X	X			
Instagram			X	X	X			X						
Facebook		X	X				X							
Twitter	X	X			X		X	X	X					

#### 4.2.4.2 Controls preprocessing

Based on our observation that the sentences related to users' choices and options are stated throughout the policy text, we used the text files saved in the original preprocessing step. We also created a list of terms describing the information or practices users can control. These terms are used to be matched to terms appearing in each service provider's policy and then to determine which control options are included in the policies.

#### 4.2.4.3 Controls matching-based extraction algorithm

We developed the controls matching-based extraction algorithm to locate the information and practices that the service providers provide to users to stop sharing or opt out of. The algorithm uses the text file (f) of the privacy policy's full text, created during original preprocessing, and a list of control terms (l) as the inputs and generates a list of control terms (c) as the output that appears in the original privacy policy text file (f). Each time a term from

the input list (l) is matched to a paragraph (p) it is added to the list (c). A match is determined in three steps. A paragraph (p) must contain a term from the input list (l). The same paragraph must also contain a “control keyword.” Finally, it must be determined that the term is not an exception. The term “SMS” is considered an exception in the controls extraction because it appears in the middle of other irrelevant terms. For example, it can be found in the following statement: “Using the choice mechanisms these programs make available may help you see advertising that is relevant to you or help you avoid seeing interest based advertising generally.”

<b>Algorithm 3: Controls Matching-Based Extraction Algorithm</b>
<b>Input:</b> Privacy policy text file (f) A list of control terms (l)
<b>Output:</b> A list of control terms covered in providers’ privacy policies (c)
<b>Begin:</b> 1 <b>for</b> each term in the list (l) i → 1 to N 2 <b>for</b> each paragraph (p) in the text j → 1 to M 3 <b>if</b> paragraph contains term 4 <b>if</b> term not an exception & p contains “control keyword” add the term to control list (c) 5 <b>else</b> skip <b>end if</b> <b>end if</b> <b>end for</b> <b>end for</b> 6. <b>return</b> list of control terms (c)
<b>End Begin</b>

#### 4.2.5 Sharing extraction

In this section, we focus on extracting information about the party with whom the users’ information is shared.

#### 4.2.5.1 Analysis process

The training dataset was analyzed to determine how sharing practices are described in service providers' privacy policies. The following observations were made during the analysis:

1. Sharing practices are described in a specific section of the policy.
2. The text that describes sharing practices often contains the two words "share" and "with." For example, the Mayo Clinic's privacy policy dictates that users' information may be shared with third parties as in the following statement: "We may *share* your personally identifiable information *with third parties* who we have engaged to help us provide the services."
3. Unlike personal information and control extraction, there are no common words or terms to describe the party with whom users' information shared. For example, in Visa's privacy policy it states the following: "We may *share* your Personal Information *with other companies that are owned by Visa as allowed by law.*" This shows that Visa uses terminologies different from Mayo Clinic's terminologies as has been shown to describe the party with whom user's information is shared.

Due to this diversity of terms, it is difficult to list the parties from the policies.

4. In the text, the party with whom the users' information is shared most commonly appears as a noun phrase after the word "with," as can be seen in the Mayo Clinic's privacy policy statement: "We may *share* your personally identifiable information *with third parties* who we have engaged to help us provide the services" and in Twitter's

privacy policy statement, “We *share* your Payment Information *with* **payment services providers** to process payments.”

Based on the fourth observation, we concluded that the texts describing the parties are noun phrases appearing after the word “with.” Thus, a part of speech technique would be helpful in automatically determining the parties and processing their extraction.

We used a Stanford POS tagger to tag the text and then find the common characteristics of party’s phrases as written in the privacy policies based on their part of speech. Figure 10 shows an example of a sentence tagged with Stanford POS tagger.

*Figure 10: Tagged sentence*

We/PRP also/RB may/MD share/VB your/PRP\$ information/NN with/IN **third-party/JJ organizations/NNS** that/WDT help/VBP us/PRP provide/VB the/DT Service/NNP to/TO you/PRP./.

While analyzing several sentences, we found that the parties’ noun phrases are tagged with several common tags and end when a specific tag is reached, defined as the “stop tag.” We define the “stop tag” as a part of speech tag that appears after a party noun phrase that indicates the end of phrases. Table 23 shows part of our analysis.

Table 23: Sample of analysis results

	Tagged sentence	Parties tags	Stop word	Stop word tag
1	We/PRP may/MD share/VB your/PRP\$ personally/RB identifiable/JJ information/NN with/IN <b>third/JJ parties/NNS</b> <u>who/WP</u> we/PRP have/VBP engaged/VBN to/TO help/VB us/PRP provide/VB the/DT services/NNS ./.	JJ NNS	who	WP
2	NET-A-PORTER/NNP may/MD also/RB share/VB your/PRP\$ data/NNS with/IN <b>third/JJ party/NN business/NN partners/NNS</b> <u>to/TO</u> provide/VB you/PRP with/IN targeted/JJ advertising/NN ./.	JJ NN NN NNS	to	TO
3	We/PRP may/MD share/VB your/PRP\$ Personal/JJ Information/NN with/IN <b>companies/NNS and/CC vendors/NNS</b> <u>that/WDT</u> help/VBP us/PRP to/TO operate/VB our/PRP\$ business/NN ./.	NNS CC NNS	that	WDT
4	We/PRP also/RB may/MD share/VB aggregate/JJ ./, non-personal/JJ information/NN about/IN website/NN usage/NN with/IN <b>unaffiliated/JJ third/JJ parties/NNS</b> ./.	JJ JJ NNS	end the sentence	-

#### 4.2.5.2 Sharing preprocessing

The data sharing section of each service provider’s policy was observed to be written in a separate section and therefore saved in a text file format.

#### 4.2.5.3 Sharing Extraction Algorithm

A sharing extraction algorithm was developed to extract information related to the parties with whom service providers share their users’ information as stated in their policies. The algorithm uses the text file (f) of the privacy policy’s data sharing section, created during sharing preprocessing, and generates a list of noun phrases (n) related to parties as the output. These noun phrases are determined through the following steps. First, the text is split into sentences (s). Only sentences that contain “share” and “with” are added to a list of sentences (l). Then, each word of each sentence in the list (l) is tagged according to its part of speech. Finally, after each sentence is tagged, the words of the noun phrases that occur directly after “with” and after a stop tag are added to the list (n).

<b>Algorithm 4:</b> Sharing Extraction Algorithm
<b>Input:</b> Privacy policy text file (f)
<b>Output:</b> A list of parties' noun phrases (n)
<b>Begin:</b> 1 <b>for</b> split the paragraph into sentences (s) $i \rightarrow 1$ to N 2 <b>if</b> (s) contains "share" and "with" add (s) to the list (l) 3 <b>else</b> skip <b>end if</b> <b>end for</b> 4 <b>for</b> each sentence in (l) $i \rightarrow 1$ to M tag the sentence 5 <b>for</b> each word after with 6 <b>if</b> tag of the word not stop tag add it to the list (n) <b>end if</b> <b>end for</b> <b>end for</b> 7. <b>return</b> list of noun phrases (n)
<b>End Begin</b>

### 4.3 Results and Evaluation

#### 4.3.1 Personal information evaluation

##### 4.3.1.1 Personal information classification

In addition to our proposed approach, we also studied a classification approach which classifies the noun phrases into personal and non-personal information. We performed several strategies during the evaluation process.

We performed an initial study by applying several classifiers including Random Forest, J48, Ibk, SMO, Logistics, and Naïve Bayes using the WEKA open-source implementation with default parameters. The initial evaluation results indicate that the Random Forest classifier provides the best accuracy results.

Considering that the Random Forest classifier showed the best results, it was used to perform further evaluation experiments. It was used to classify different dimensions of the datasets. Singular value decomposition dimension reduction method was applied to the dataset using MATLAB tool to produce several versions of the dataset with different dimensions. Table 24 shows the evaluation results of the different dataset dimensions. The results indicate that the dataset reduced to six dimensions gives the best evaluation results compared other dimensions, but the results were still not satisfactory. Thus, the dataset reduced to six dimensions was used in subsequent experiments trying to improve classifier accuracy.

*Table 24: Random Forest evaluation results using different dimensions*

# Dimensions	Precision	Recall	F-Score	ROC
555	1	0.04	0.08	0.68
500	1	0.13	0.23	0.7
400	1	0.15	0.26	0.78
300	1	0.17	0.3	0.83
200	0.88	0.17	0.29	0.87
100	0.86	0.28	0.43	0.89
50	0.84	0.35	0.5	0.92
25	0.73	0.48	0.58	0.93
15	0.75	0.51	0.63	0.89
10	0.73	0.48	0.58	0.87
<b>6</b>	<b>0.75</b>	<b>0.6</b>	<b>0.64</b>	<b>0.91</b>

Due to the imbalance of the dataset in which the number of personal information instances is too low compared to the number of non-personal information instances, we applied and tested down-sampling and up-sampling techniques using the six-dimension dataset. Down-sampling is the technique of removing some data from the majority class to balance the dataset. We tried three different combinations for down-sampling: (a) the removal

of non-personal information instances, making both classes' instances equal; (b) the removal of non-personal information instances, making the number of non-personal information instances equals double the number of personal information instances; and (c) the removal of non-personal information instances, making the number of non-personal information instances equal triple the number of personal information instances.

Up-sampling is the technique of duplicating instances of the minority class to balance it with the number of instances in the majority class. This technique was used during the evaluation process by duplicating the number of personal information instances to make them equal to the number of non-personal information instances. Table 25 shows the evaluation results of applying the techniques with different classifiers.

Table 25: Sampling evaluation results. (p) is personal information class, (np) is non-personal information class.

	Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall		Precision Recall	
	Random forest		J48		Ibk		SMO		Logistic		Naïve Bayes	
Original Data	0.75	0.6	0.56	0.6	0.69	0.64	1	0.13	0.84	0.46	0.62	0.44
UpSampling-50%(p) 50% (np)	0.76	0.66	0.55	0.6	0.69	0.64	0.55	0.55	0.3	0.71	0.42	0.55
DownSampling-50%(p) 50% (np)	0.29	0.91	0.24	0.93	0.35	0.88	0.69	0.51	0.23	0.97	0.2	0.57
DownSampling-33%(p) 66% (np)	0.49	0.86	0.33	0.77	0.49	0.84	0.82	0.42	0.41	0.66	0.22	0.53
DownSampling-25%(p) 75% (np)	0.58	0.84	0.33	0.75	0.49	0.84	0.81	0.4	0.46	0.57	0.36	0.48

The table results show that up-sampling and down-sampling techniques improve the recall values of different classifiers; however, they reduce the precision values. This means that various numbers of noun phrases were labeled as personal information, but they were not.

#### 4.3.1.2 Personal information rule-based extraction

We used our testing dataset to validate the rule-based extraction. We created the five components of the rule table and tried to match the sentences from the policies of the testing

dataset into the rules created using the training dataset. We found several unmatched examples where a sentence was organized differently compared to sentences in the training dataset. For example, in Barney’s privacy policy “Whenever you shop with Barneys New York we obtain from you the information we need to complete your transaction. This may include your name, address, telephone number, e-mail address, and credit card number.” We were unable to detect the collected personal information as some components were in one sentence and the personal data component was in another sentence, which does not match any of our previously defined rules. Table 26 shows the evaluation results of the testing dataset.

*Table 26: Testing dataset evaluation results- personal information rule-based extraction*

Service Provider	Retrieved	relevant	R&R	Precision	Recall	F1 score
Kemper	9	10	9	1	0.9	0.95
Hotels	14	15	14	1	0.93	0.97
SSA	0	0	0	-	-	-
Mass	4	6	4	1	0.67	0.80
Priceline	4	12	4	1	0.33	0.50
Shopbop	14	15	14	1	0.93	0.97
WesternU	12	15	5	0.42	0.33	0.37
Humana	2	1	1	0.5	1	0.67
Chase	3	3	3	1	1	1.00
Barneys	0	1	0	-	0	-
LinkedIn	5	4	4	0.8	1	0.89
Tumblr	7	14	7	1	0.5	0.67
Total	<b>74</b>	<b>96</b>	<b>65</b>	<b>0.88</b>	<b>0.68</b>	<b>0.76</b>

#### 4.3.1.3 Personal information matching-based extraction

During its development, the algorithm was tested iteratively with the training set. We call this step a validation process. The fact that four domains were included in the study helped during the validation process because we could test a wide range of cases, allowing us to exclude many exceptions. Also, we could verify the completeness of the list of personal information terms in several ways. First, we used it to ensure that the list of terms included the

personal information collected by all the websites. For example, it is probable that the word “religion” will appear in social network privacy policies, but it would not be expected to appear in the privacy policies of financial domains. Second, we used it to ensure that our list contains the most basic versions of terms to allow for the greatest possible matching with all potential forms of any one concept in the privacy policy text. For instance, the term “birth” is used in the list (I) as a “search term” because it can be matched to the terms “birthday,” “date of birth,” and “birthdate,” thereby gathering the most possible data from the privacy policy. Third, we included abbreviations in our list to ensure that matches were made even when service providers choose not to use full forms of keywords. For example, searching for “DOB,” the abbreviation for “date of birth,” helps ensure that all applicable personal information related to a user’s date of birth is extracted from the privacy policy regardless of a service provider’s choice of terminology.

We included four test cases during the validation process and the evaluation of the training dataset results: match without action words, match without action words—distinct, match with action words, and match with action words—distinct. The reasons for including these four cases include testing whether the consideration of the action words’ appearance in the information collection section text improves the algorithm’s accuracy. We evaluated the results with and without the implementation of action words. Second, because our main goal was to extract the personal information collected by the service providers as stated in their privacy policies regardless of how many times the information is stated, we performed the evaluation based on the results of all extracted personal information as well as the extraction of

distinct results only. For instance, if the term “location” appears in two paragraphs of the text and our algorithm were able only to capture one occurrence, then the accuracy will be 100% instead of 50%.

The output list of personal information terms may contain either “relevant” or “irrelevant” terms. “Relevant” terms are personal information terms related to a service provider’s collection practices. For example, the term “social security” appearing in CIGNA’s privacy policy statement, “We collect personal information that you voluntarily provide through our Services, including: Name, address, and birthdate; Other contact information such as e-mail address and/or text address; Financial and health information; Credit or debit card number; and Social Security or similar national ID number; Geolocation information; and Social media account IDs.” “Irrelevant” terms are personal information terms that are retrieved by the algorithm but are not related to the collection practices. This would include the terms “name” and “username” as they appear in Instagram’s privacy policy statement: “If you choose to find your friends (iii) through a search of **names** or **usernames** on Instagram then simply type a name to search and we will perform a search on our Service.”

In addition, “missing” terms are the information related to collection practice but are not extracted by the algorithm. For example, the term “financial information” is missed when matched to CIGNA’s following statement, “We collect personal information that you voluntarily provide through our Services, including: Name, address, and birthdate; Other contact information such as e-mail address and/or text address; **Financial** and health information;

Credit or debit card number; and Social Security or similar national ID number; Geolocation information; and Social media account IDs.”

Tables 27 – 30 show the results of the training dataset, and Table 31 shows the precision, recall, and F<sub>1</sub> score results of the four cases.

Table 27: Results of all output – no action words

Service providers	Relevant	Retrieved	R&R
Quest Diagnostics	6	6	6
CIGNA	10	9	9
Mayo	0	0	0
Visa	8	8	8
CitiBank	6	6	6
BoA	7	7	7
Bloomingdale's	10	9	9
Neiman Marcus	12	12	12
Net-A-Porter	11	12	11
Instagram	10	9	9
Facebook	8	9	8
Twitter	20	22	20
<b>Total</b>	<b>108</b>	<b>109</b>	<b>105</b>

Table 28: Results of distinct outputs – no action words

Service providers	Relevant	Retrieved	R&R
Quest Diagnostics	6	6	6
CIGNA	10	9	9
Mayo	0	0	0
Visa	8	9	8
CitiBank	6	7	6
BoA	7	7	7
Bloomingdale's	11	10	10
Neiman Marcus	13	14	13
Net-A-Porter	14	16	14
Instagram	23	23	22
Facebook	11	13	11
Twitter	37	41	37
<b>Total</b>	<b>146</b>	<b>155</b>	<b>143</b>

Table 29: Results of distinct outputs – action words

Service providers	Relevant	Retrieved	R&R
Quest Diagnostics	6	6	6
CIGNA	10	9	9
Mayo	0	0	0
Visa	8	8	8
CitiBank	6	6	6
BoA	7	7	7
Bloomingdale's	10	9	9
Neiman Marcus	12	12	12
Net-A-Porter	11	12	11
Instagram	10	10	9
Facebook	7	8	7
Twitter	20	22	20
<b>Total</b>	<b>107</b>	<b>109</b>	<b>104</b>

Table 30: Results of all output – action words

Service providers	Relevant	Retrieved	R&R
Quest Diagnostics	6	6	6
CIGNA	10	9	9
Mayo	0	0	0
Visa	8	9	8
CitiBank	6	7	6
BoA	7	7	7
Bloomingdale's	10	9	9
Neiman Marcus	13	13	13
Net-A-Porter	14	16	14
Instagram	14	18	13
Facebook	10	11	10
Twitter	37	41	37
<b>Total</b>	<b>135</b>	<b>146</b>	<b>132</b>

Table 31: Training dataset evaluation results- personal information matching-based extraction

Case	Precision	Recall	F1 score
Match w/o action words	0.98	0.92	0.95
Match w/o action words- distinct	0.96	0.97	0.97
Match with action words	0.90	0.98	0.94
Match with action words- distinct	0.95	0.97	0.96

#### 4.3.1.3.1 Testing dataset evaluation

After reaching a reasonable result, with an  $F_1$  score above 90%, during the development process by modifying the list and editing the algorithm, we started to evaluate the algorithm with the testing dataset. Tables 32-35 show the results of the testing.

Table 32: Results of distinct outputs – no action words

Service Provider	Retrieved	relevant	R&R
Kemper	9	10	9
Hotels	14	15	14
SSA	4	3	3
Mass	4	6	4
Priceline	8	8	8
Shopbop	8	9	8
WesternU	10	12	10
Humana	9	9	9
Chase	12	12	12
Barneys	9	9	9
LinkedIn	16	15	15
Tumblr	16	14	14
<b>Total</b>	<b>119</b>	<b>122</b>	<b>115</b>

Table 33: Results of distinct outputs – action words

Service Provider	Retrieved	Relevant	R&R
Kemper	9	10	9
Hotels	14	15	14
SSA	4	3	3
Mass	4	6	4
Priceline	8	8	8
Shopbop	8	9	8
WesternU	10	12	10
Humana	9	9	9
Chase	12	12	12
Barneys	9	9	9
LinkedIn	16	15	15
Tumblr	16	14	14
<b>Total</b>	<b>119</b>	<b>122</b>	<b>115</b>

Table 34: Results of all outputs – no action words

Service Provider	Retrieved	relevant	R&R
Kemper	10	11	10
Hotels	18	16	15
SSA	4	3	3
Mass	4	6	4
Priceline	12	11	11
Shopbop	16	15	14
WesternU	10	12	10
Humana	15	14	14
Chase	13	13	13
Barneys	9	9	9
LinkedIn	42	37	37
Tumblr	27	23	23
<b>Total</b>	<b>180</b>	<b>170</b>	<b>163</b>

Table 35: Results of all outputs – action words

Service Provider	Retrieved	Relevant	R&R
Kemper	10	11	10
Hotels	18	16	15
SSA	4	3	3
Mass	4	6	4
Priceline	12	11	11
Shopbop	16	15	14
WesternU	10	12	10
Humana	9	9	9
Chase	12	12	12
Barneys	9	9	9
LinkedIn	41	37	35
Tumblr	27	23	23
<b>Total</b>	<b>172</b>	<b>164</b>	<b>155</b>

An example of a missing term when examining LinkedIn’s policy is “IP address.” It is missed because it is stated, “We also receive the internet protocol (“IP”) address of your computer or the proxy server that you use to access the web.” The presence of the parentheses and quotation marks makes the term different; thus, the algorithm was not able to detect that. An example of irrelevant extraction is the term “email” in the following LinkedIn statement: “Another example are software tools that allow you to see our and other public information about the people you **email** or meet with...” Table 36 shows the evaluation results of the testing dataset.

*Table 36: Testing dataset evaluation results- personal information matching-based extraction*

Case	Precision	Recall	F1 score
Match w/o action words	0.91	0.96	0.93
Match w/o action words- distinct	<b>0.97</b>	<b>0.94</b>	<b>0.95</b>
Match with action words	0.90	0.95	0.92
Match with action words- distinct	<b>0.97</b>	<b>0.94</b>	<b>0.95</b>

#### 4.3.2 Controls extraction evaluation

We tested the algorithm iteratively during the validation process using the training dataset to assess the handling of exceptions and the coverage of list terms. In this section, we included the following four evaluation cases to evaluate the output results: match without control keywords, match without control keywords—distinct, match with control keywords, and match with control keywords—distinct.

Tables 37 – 40 show the results of the extraction of control terms from the training dataset. Table 41 shows the precision, recall, and  $F_1$  score results of the four cases.

Table 37: Results of distinct outputs – no control keywords

Service provider	Retrieved	Relevant	R &R
Quest Diagnostics	2	1	1
CIGNA	10	6	6
Mayo	10	7	7
Visa	8	7	7
CitiBank	8	7	7
BoA	10	7	7
Bloomingdale's	9	7	7
Neiman Marcus	10	10	10
Net-A-Porter	9	6	6
Instagram	8	5	5
Facebook	8	5	5
Twitter	9	6	6
<b>Total</b>	<b>101</b>	<b>74</b>	<b>74</b>

Table 38: Results of all outputs – no control keywords

Service provider	Retrieved	Relevant	R &R
Quest Diagnostics	2	1	1
CIGNA	6	6	6
Mayo	18	15	15
Visa	10	8	8
CitiBank	11	11	11
BoA	19	19	19
Bloomingdale's	15	14	14
Neiman Marcus	12	11	11
Net-A-Porter	9	7	7
Instagram	12	10	10
Facebook	6	5	5
Twitter	14	9	9
<b>Total</b>	<b>134</b>	<b>116</b>	<b>116</b>

Table 39: Results of distinct outputs – control keywords

Service provider	Retrieved	Relevant	R &R
Quest Diagnostics	1	1	1
CIGNA	6	6	6
Mayo	6	7	6
Visa	5	7	5
CitiBank	7	7	7
BoA	7	7	7
Bloomingdale's	7	7	7
Neiman Marcus	9	10	9
Net-A-Porter	6	6	6
Instagram	4	5	4
Facebook	4	5	4
Twitter	6	6	6
<b>Total</b>	<b>68</b>	<b>74</b>	<b>68</b>

Table 40: Results of all outputs – control keywords

Service provider	Retrieved	Relevant	R &R
Quest Diagnostics	5	1	1
CIGNA	30	6	6
Mayo	47	15	15
Visa	23	8	8
CitiBank	44	11	11
BoA	65	19	19
Bloomingdale's	35	14	14
Neiman Marcus	42	11	11
Net-A-Porter	27	7	7
Instagram	29	10	10
Facebook	28	5	5
Twitter	30	9	9
<b>Total</b>	<b>405</b>	<b>116</b>	<b>116</b>

Table 41: Training dataset evaluation results- controls extraction

Case	Precision	Recall	F <sub>1</sub> score
Match w/o control keywords	0.29	1	0.45
Match w/o control keywords- distinct	0.73	1	0.85
Match with control keywords	0.87	1	0.93
Match with control keywords- distinct	1	0.92	0.96

#### 4.3.2.1 Evaluation of testing dataset

Achieving an  $F_1$  score above 90% for the cases involving matching with control keywords is a reasonable result. Once that  $F_1$  score was achieved, we terminated the validation process of the training dataset and started applying the algorithm to the testing dataset. Tables 42 – 45 show the results of applying the control matching-based extraction on the testing dataset.

Table 42: Results of distinct outputs – no control keywords

Service Provider	Retrieved	relevant	R&R
Kemper	1	1	1
Hotels	8	7	7
SSA	3	1	1
Mass	5	3	3
Priceline	8	6	6
Shopbop	9	8	8
WesternU	5	0	0
Humana	7	5	5
Chase	9	4	4
Barneys	4	3	3
LinkedIn	10	9	8
Tumblr	5	3	3
<b>Total</b>	<b>74</b>	<b>50</b>	<b>49</b>

Table 43: Results of all outputs – no control keywords

Service Provider	Retrieved	Relevant	R&R
Kemper	2	2	2
Hotels	40	11	11
SSA	10	2	2
Mass	18	4	4
Priceline	43	19	19
Shopbop	22	14	14
WesternU	14	0	0
Humana	22	8	8
Chase	30	8	8
Barneys	10	3	3
LinkedIn	90	22	21
Tumblr	15	7	7
<b>Total</b>	<b>316</b>	<b>100</b>	<b>99</b>

Table 44: Results of distinct outputs – control keywords

Service Provider	Retrieved	Relevant	R&R
Kemper	1	1	1
Hotels	7	8	7
SSA	2	1	1
Mass	3	4	3
Priceline	5	5	5
Shopbop	5	6	5
WesternU	0	0	0
Humana	5	5	5
Chase	4	4	4
Barneys	3	3	3
LinkedIn	8	9	8
Tumblr	2	2	2
<b>Total</b>	<b>45</b>	<b>48</b>	<b>44</b>

Table 45: Results of all outputs – control keywords

Service Provider	Retrieved	relevant	R&R
Kemper	1	2	1
Hotels	16	12	11
SSA	2	1	1
Mass	3	4	3
Priceline	13	13	13
Shopbop	8	6	5
WesternU	0	0	0
Humana	8	8	8
Chase	8	8	8
Barneys	3	3	3
LinkedIn	27	22	21
Tumblr	5	4	4
<b>Total</b>	<b>94</b>	<b>83</b>	<b>78</b>

An example of a missing term when examining LinkedIn’s policy is “survey.” It was missed because none of the policies in the training dataset has the same term when describing users’ options, and so it was not included in the control list. An example of irrelevant extraction

is the term “location” in the following Humana statement: “Geo-location data includes zip code entered by the user for use within the Provider/Pharmacy search function.” Table 46 shows the evaluation results of the testing dataset.

Table 46: Testing dataset evaluation results- controls extraction

Case	Precision	Recall	F1 score
Match w/o action words	0.31	0.99	0.48
Match w/o action words- distinct	0.66	0.98	0.79
Match with action words	0.83	0.95	0.89
<b>Match with action words- distinct</b>	<b>0.98</b>	<b>0.92</b>	<b>0.95</b>

#### 4.3.3 Sharing extraction evaluation

We tested the algorithm repeatedly throughout the validation process to verify that all possible cases of stop tags were included. Table 47 shows the results and evaluation of the training dataset.

Table 47: Training dataset evaluation results- sharing extraction

Service provider	Retrieved	Relevant	R &R	Precision	Recall	F1 score
Quest Diagnostics	1	1	1	1	1	1
CIGNA	7	7	5	0.71	0.71	0.71
Mayo	1	1	1	1	1	1
Visa	7	5	5	0.71	1	0.83
CitiBank	4	3	3	0.75	1	0.86
BoA	4	3	3	0.75	1	0.86
Bloomingdale's	7	6	6	0.86	1	0.92
Neiman Marcus	3	2	2	0.67	1	0.8
Net-A-Porter	7	8	6	0.86	0.75	0.8
Instagram	6	5	5	0.83	1	0.91
Facebook	7	7	7	1	1	1
Twitter	6	6	5	0.83	0.83	0.83
<b>Total</b>	<b>60</b>	<b>54</b>	<b>49</b>	<b>0.82</b>	<b>0.91</b>	<b>0.86</b>

#### 4.3.3.1 Evaluation of the testing dataset

Table 48 shows evaluation results of testing dataset.

Table 48: Testing dataset evaluation results- sharing extraction

Service provider	Retrieved	Relevant	R &R	Precision	Recall	F <sub>1</sub> score
Kemper	9	10	9	1	0.9	0.95
Hotels	13	11	11	0.85	1	0.92
SSA	2	3	2	1	0.67	0.8
Mass	1	1	1	1	1	1
Priceline	2	3	2	1	0.67	0.8
Shopbop	2	2	2	1	1	1
WesternU	7	5	5	0.71	1	0.83
Tumblr	12	11	11	0.92	1	0.96
Humana	2	1	1	0.5	1	0.67
Chase	3	3	3	1	1	1
Barneys	0	1	0	-	0	-
LinedIn	5	4	4	0.8	1	0.89
<b>Total</b>	<b>58</b>	<b>55</b>	<b>51</b>	<b>0.87931</b>	<b>0.93</b>	<b>0.90</b>

#### 4.4 Discussion and conclusion

Although different experiments were performed and several techniques were applied to improve the personal information classification-based and rule-based accuracy, the results were still not satisfactory. Thus, we introduced another approach, the matching-based extraction approach.

We developed three algorithms to extract data related to information collection, sharing, and control practices. The overall accuracy represented by an F<sub>1</sub> score of personal information extraction is above 92%. The implementation of action words in the personal information extraction algorithm has no notable improvement on the algorithm's accuracy because it only improves from 92% to 93%. The slight improvement occurs because the analyzed text was from the data collection section. The appearance of personal information in

that section would be expected to be relevant to data collection practices. The results also show that the accuracy increased slightly from 92% to 95% when considering distinct information. When the algorithm was tested with the testing dataset, during the evaluation, we did not find any missing personal information with the reason that it was not included in the list of terms. This indicates the completeness of the list and applicable nature of the solution.

In the extraction of control-related terminology, the accuracy of the algorithm without the inclusion of control keywords is very low, 48% and 79%, compared to the accuracy of the algorithm with the inclusion of control keywords, 89% and 95%. This increased accuracy occurs because the terminology related to control practices can appear anywhere in the text, not just in one specific section. The implementation of control keywords adds context to the extraction, improving the accuracy. The test cases that include distinct information also show improvement in algorithm accuracy. The accuracy increases from 48% to 79% without including control keywords and, when including control keywords, from 89% to 95%.

One of the limitations of the extraction system is that it depends on the accuracy of the POS tagger. Whenever the POS tagger produces incorrect results, the accuracy of the extraction system will be negatively affected (Constante et al., 2012). In our algorithm, although we used a POS tagger in the process of extracting information from sharing section of privacy policy, we did not find that POS tagging had a negative impact on the algorithm accuracy during the experiment.

One of the reasons that noun phrases referring to parties were missed is that the algorithm does not handle the coreference resolution. Coreference resolution is the task of

finding all expressions that refer to the same entity in a text. An example that explains a missed extraction due to a lack of coreference resolution is a statement in CIGNA's privacy policy, "We may share your personal information with them so they can provide those services." Here the word "them" refers to a "third party" from a previous sentence, but because the coreference is not applied, the algorithm will not be able to handle that.

Although we believe that the extraction of 90% of the information related to sharing practices is reasonable, we suggest applying coreference resolution technique to catch the missing phrases and improve the accuracy. Another suggested improvement is to apply machine learning techniques to define automatically the common text characteristics of sharing practices and the determination of stop words instead of implementing that manually.

Finally, the extraction of important policy content with a minimum accuracy of 90% means that, when this content is displayed in a way that is easy to read and comprehend, users' awareness will be improved by at least 90% compared to the users who do not read the policies due to the policy length or difficulty.

## CHAPTER 5: Privacy Policies Comparison

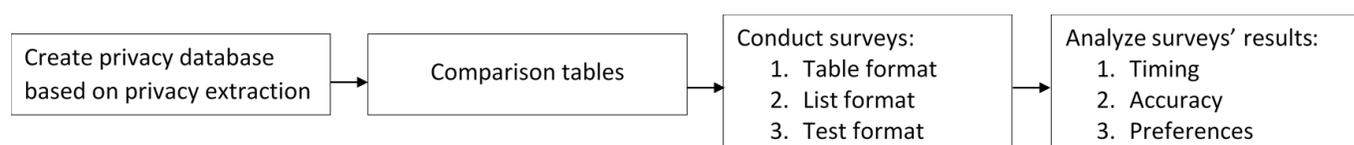
### 5.1 Introduction

Studies have indicated that reading full privacy policies is time-consuming and challenging. Users must spend a long time reading full privacy policies and have difficulty correctly answering questions about privacy practices stated in service providers' policies (Kelley et. al, 2010). We think that users will find more difficulty in answering such questions when refereeing to two service providers' privacy policies. To answer questions related to privacy practices users will need to read the first policy, remember what they read in it, read the second policy, and then compare the two policies.

Previous studies have introduced several approaches to improve privacy policy readability. Some studies have proposed standardized policy formatting to simplify the readability of a single privacy policy. To the best of our knowledge, no study has been introduced to highlight privacy policy contents to help users compare privacy practices of two or more service providers.

In this chapter, we highlight the privacy policy contents that describe privacy practices. (Users found them important or not clear in privacy concerns study results, as explained in Chapter 3.) We present contents related to information collection, sharing, and control in pairs of policies. Each policy in a pair is from the same domain as the other, and they are presented side-by-side to allow users to locate and compare this information easily between the policies. We conducted a survey to study how effective highlighting these contents would be in helping

users to gain information, reduce time to look up information, and simplify the comparison of specific information between two policies. Moreover, we examine users' privacy preferences drawn from the presented privacy practices, which will help us to rank privacy policies in the next chapter. Figure 11 shows an overall steps of privacy policies comparison.



*Figure 11 Overall steps of privacy policies comparison*

## 5.2 Privacy database

### 5.2.1 Information category

The output lists from our information extraction process were used in this chapter to conduct surveys. The specific information of each privacy practice is categorized into general levels. The categorization of the information could help the users to add context to a specific piece of information which help them find it easily. For example, categorizing “card number” into “Financial information” will help users to recognize the collected number is either related to their credit card or debit card. Personal information is classified into four categories: contact, financial, geolocation, and identity. The controls and options provided to users are classified into six categories: activity, advertisements and marketing, communication, geographical, personal information, and research. The third parties with whom users' information is shared are divided into four categories: user, affiliate, non-affiliate, and unspecified party. Parties are classified as “user” when the policy states that the information is shared with other users using

the same service. Parties are classified as “affiliate” when the word “affiliate” is used or when the policy states that the party is within the same company network or service providers.

Parties are classified as “non-affiliate” if the policy uses the word “non-affiliate” to describe the party. Finally, parties are classified as “unspecified” when they do not fall into any previous categories. Table 49 shows data categories for each privacy practice.

*Table 49: Categorization of privacy practices*

<b>Personal information</b>	<b>Controls</b>	<b>Parties</b>
Contact	Activity	User
Educational	Advertisement and Marketing	Affiliate
Financial	Communication	Non-Affiliate
Geographical	Geographical	Unspecified Party
Identity	Personal information	-
-	Research	

### 5.2.2 Database schema

To build a database of privacy practices, we first designed a schema to define tables, attributes of the tables, and relationships between the tables. Four tables were created to store information related to collected personal information: an extracted personal information table, a PI terms table, a PI terms-to-category table, and a PI category table. We created four other tables to store control-related information: an extracted controls table, a control terms table, a control terms-to-category table, and a controls category table. To store sharing-related information, four tables were created: an extracted parties table, a party phrases table, a phrases-to-category table, and a controls category table. A connection table, a policies table, was used to define policies’ ids and names. Figure 12 shows the schema design of the tables and the relationship between them.

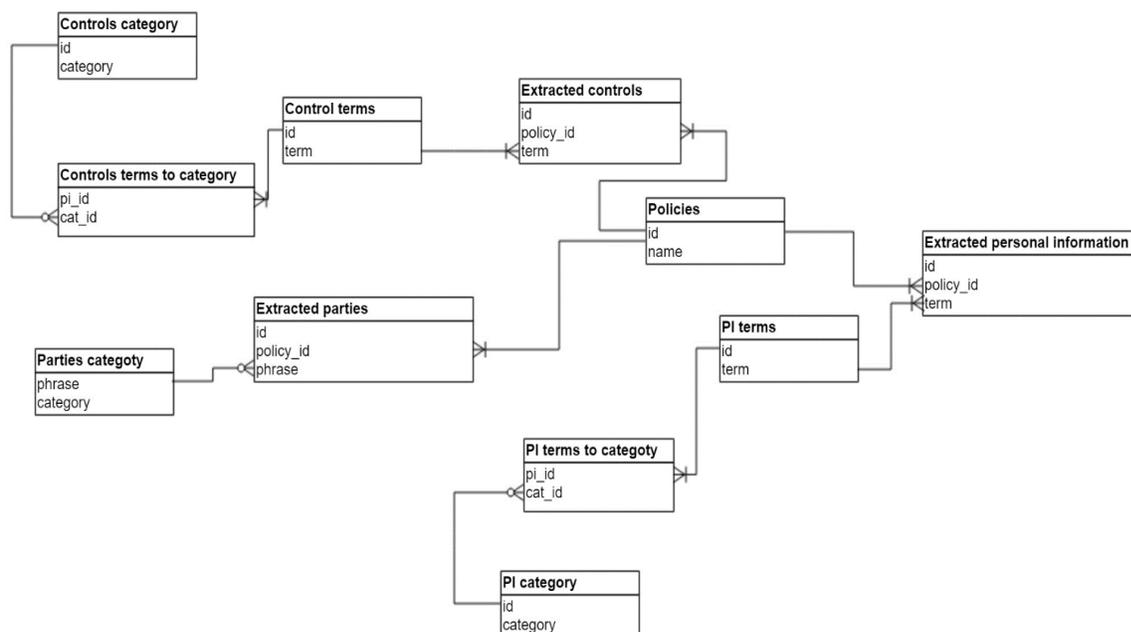


Figure 12: Schema design

### 5.3 Comparison tables

Comparison tables were designed to present information related to privacy practices of service providers. The first column in the personal information comparison table lists the categories of personal information, the second column lists potential personal information collected by service providers, and the third and fourth columns represent a pair of service providers. An “x” is used to indicate that a specific piece of information is collected by the associated service provider. An example of a personal information comparison table is shown in Table 50.

Table 50: Personal information comparison table

Personal information Category	Personal information	Company A	Company B
Contact	email address	x	x
	mailing address	x	
	phone number		x
Financial	transactions		x
Identity	age	x	
	interest	x	
	IP address	x	x
	name	x	x

In the controls comparison table, the first column shows the categories of controls or options-related terms, the second column shows the specific control-related term, and the third and fourth columns represent a pair of service providers. An “x” is used to indicate that the user can control the specific data or behavior for the associated service provider. Table 51 shows an example of a controls comparison table.

Table 51: Controls comparison table

Control Category	Control	Company A	Company B
<b>Activity</b>	cookies	x	x
<b>Advertisements and Marketing</b>	advertisements		x
	flash		x
	marketing	x	x
	offers	x	x
	promotions	x	
<b>Communication</b>	mails	x	x

In the sharing comparison table, the first column shows the categories of the parties with whom users’ information is shared, while the second and third columns represent a pair of service providers. An “x” is used to indicate the party with whom each service provider shares users’ information. Table 52 shows an example of sharing comparison table.

Table 52: Sharing comparison table

Third Party	Company A	Company B
Affiliate		x
Non-Affiliate	x	
Unspecified third party		x

## 5.4 Research questions

Our study was designed to answer the following questions:

1. Are users able to compare the privacy practices of two policies shown side-by-side in a comparison table faster than when reading each policy's full text?
2. Are users able to receive accurate information when looking at two policies shown side-by-side in a comparison table compared to when reading each policy's full-length text?
3. Which company is most preferred by the users based on its information collection, sharing, and control practices?

## 5.5 Study design

We used 20 privacy policies from four domains with five policies per domain. For the purpose of comparison, we created ten pairs of providers per domain, for a total of 40 pairs of providers across the study. For each pair, three versions of the survey were created. One version shows the privacy policies' highlighted information in tables, another version shows the privacy policies' highlighted information in lists, and the third version shows the original policies' text (40 pairs represented by tables, 40 pairs presented by lists, and 40 pairs represented by the policies' original text, totaling 120 pairing representations.) In the survey, we chose to anonymize the service providers' names to avoid bias based on brand effect and previous knowledge. Table 53 shows each service provider's anonymized name.

Table 53: Service providers anonymized name

Domain		Service Provider		Anonymized name
1	Health care	1	CIGNA	Company A
		2	Mayo Clinic	Company B
		3	Quest Diagnostics	Company C
		4	Humana	Company D
		5	Massachusetts General Hospital	Company E
2	Financial	6	Bank of America	Company F
		7	Citibank	Company G
		8	Visa	Company H
		9	JPMorgan Chase Bank	Company I
		10	Western Union	Company J
3	Social networking	11	Facebook	Company K
		12	Twitter	Company L
		13	Instagram	Company M
		14	LinkedIn	Company N
		15	Tumblr	Company O
4	Shopping	16	Bloomingdale's	Company P
		17	Neiman Marcus	Company Q
		18	Net-A-Porter	Company R
		19	Barney NY	Company S
		20	Shopbop	Company T

We used Quicksurveys.com to design the surveys. Each survey is divided into several sections.

1. *Demographics*: Information was collected about the gender, age range, and educational level of participants.
2. *Privacy concerns*: We asked participants about their privacy concerns related to the privacy of their personal information on the Internet. They were asked to rate their concerns according to one of the five levels: extremely concerned, very concerned, moderately concerned, slightly concerned, and not at all concerned.

3. *Information collection:* Participants were shown either a personal information comparison table, two lists of highlighted personal information, or two texts of privacy policies related to information collection practices. Participants were then asked which company collects specific information and which company collects more information. These questions were asked to test the hypothesis that participants using highlighted information in the tables and lists will find it easier to find the correct answers than participants using a text version of a privacy policy. Also, in this section participants were asked whether they would prefer to deal with one company rather than the other based on the presented information related to the companies' collection practices.
4. *Controls:* Participants were shown either a controls comparison table, two lists of provided controls, or a full-text version of two companies' policies text. Participants were then asked which company allows its users to control specific data or behavior and which company gives their users' more controls over information collection and sharing. Finally, based on the control-related practices of the two companies surveyed, participants were asked if they would prefer one company to the other.
5. *Sharing:* Participants were shown either a sharing comparison table, two lists of parties with whom users' information is shared, or part of a text policy containing the parties with whom two companies share users' information. Participants were then asked which company shares users' information with a specific party and which of the two companies shares users' information with more parties. Finally, based on sharing

practices, participants were asked to choose whether they would prefer one company to the other.

6. *General preference and comparison simplicity/easiness*: In this section, participants were asked if they would prefer one company to the other based on their general privacy practices. Then they were asked if they found it easy to compare two policies and answer the questions.

#### 5.5.1 Survey sample

**Worker ID:**

**Gender:**

- Male
- Female

**Age range:**

- 18-25
- 26-35
- 36-45
- 46+

**Highest level of education:**

- Some high school
- High school diploma
- Associates/ professional degree
- Some college
- Bachelor's degree
- Master's degree
- PhD
- Other postgraduate degree

**In general, regarding the privacy of your personal information on the Internet, which of the following best describes your current level of concern?**

- Extremely concerned
- Very concerned
- Moderately concerned
- Slightly concerned
- Not at all concerned

### INFORMATION COLLECTION

The following table represents the content of the information collection sections of the privacy practices of two **financial institutions**. The first column displays a list of data categories. The second column shows a list of personal information that falls within each category. The “x” symbol under each institution indicates that specific personal information is collected by the associated institution. Please refer to the table to answer the following questions.

Personal information Category	Personal information	Company A	Company B
Contact	email address	x	x
	mailing address	x	
	phone number		x
Financial	transactions		x
Identity	age	x	
	interest	x	
	IP address	x	x
	name	x	x

**Which company collects the user's phone number?**

- Company A
- Company B
- Both companies
- Neither company

**Which company collects more personal information about its users?**

- Company A
- Company B
- No difference
- Cannot determine from the table

**Assuming both companies offer the exact same services, based on their information collection practice which company would you prefer to use?**

- Company A
- Company B
- I do not have a preference

### CONTROLS

The following table represents information related to user controls and options for specific data and practices as stated in the privacy policies of two **financial institutions**. The first column displays a list of categories in which users can have direct control over the use of their personal data. The second column lists specific types of data or action within each category. The “x” symbol indicates that users have control over specified data or action with the associated institution.

Please refer to the table to answer the following questions.

Control Category	Control	Company A	Company B
<b>Activity</b>	cookies	x	x
<b>Advertisements and Marketing</b>	advertisements		x
	flash		x
	marketing	x	x
	offers	x	x
	promotions	x	
<b>Communication</b>	mails	x	x

**Which company allows their users to opt-out of receiving marketing materials?**

- Company A
- Company B
- Both companies
- Neither company

**Which company gives users more control over the use of their personal information?**

- Company A
- Company B
- No difference
- Cannot determine from the table

**If the two companies offer the exact same services, based on the control options provided by each company, which service provider would you prefer to use?**

- Company A
- Company B
- No preference

**SHARING**

The following table represents the content related to the sharing practices of two **financial institutions** as stated in their privacy policies. The first column displays the categories to which third parties may belong. The “x” symbol under each institution indicates that users’ information may be shared with the specified party of the associated institution.

Please refer to the table to answer the following questions.

<b>Third Party</b>	<b>Company A</b>	<b>Company A</b>
Affiliate	x	x
Non-affiliate		x
Unspecified third party	x	x

**Which company shares users’ information with Non-affiliate third parties?**

- Company A
- Company B
- Both companies
- Neither company

**Which company shares users' information with more parties?**

- Company A
- Company B
- No difference
- Cannot determine from the table

**Assuming both companies offer the exact same services, based on the information sharing practices, which company would you prefer to use?**

- Company A
- Company B
- No preference

**Based on the general privacy practices of the two companies, which one would you prefer to use?**

- Company A
- Company B
- No preference

**How easy was it to compare the two companies’ privacy practices and find answers to the given questions?**

- Very easy
- Easy
- Moderate
- Hard
- Very Hard

We published the survey through the crowdsourcing tool Amazon Mechanical Turk (AMT) to target diverse workers in which participants are paid to complete a task. Participants were paid at a rate of \$0.75 per HIT. We collected a total of 600 participants. 74% were males and 26% were females. Table 54 show the age distribution of the participants.

*Table 54: Age distribution of the survey participants*

Age range	18-25	26-30	31-35	36-40	40+	Total
Number of participants	144	312	72	72	144	600
Percentage	24%	52%	12%	12%	24%	100%

All participants were American as we required that as one of the HITs requirements. We also a worker to be master worker to participant in our study. Hiring master workers guarantees to get reliable results as workers with this qualification have good performance.

A HIT was created as a link for each survey. Every HIT contains an introduction of the survey briefly explaining the survey contents, the purpose of the study, and the time required to complete the survey. Each HIT also contained a link to a survey with a box for a survey code to be entered for verification and approval.

Our study used a between-subject design where different participants answer different versions of the survey. This study design helps avoid learning effects. For questions related to the sections on information collection and controls, participants were asked the exact same questions of the three survey versions of a specific pair of companies. However, for questions related to sharing practices, participants were asked different questions. Questions asked of participants looking at the comparison table and list versions were asked about a specific

party's category, while participants reading the text version were asked about parties exactly as mentioned in the text.

We published each version of the survey separately and required 5 participants per HIT. On average participants spent an average of 3:27 minutes to complete the survey related to the table; 3:17 minutes to complete the survey related to the list; and 5:44 minutes to complete the survey for the text version.

## 5.7 Results

In this section, we summarize timing results, describe accuracy results, and conclude by providing the results of user preferences in regards to the information collection, sharing, and control practices of different companies.

### 5.7.1 Timing results

We calculated the average time the participants took to complete each pair of surveys in a specific domain. Then we averaged the time of that specific domain. This process repeated for all other domains for all three versions of the survey. Tables 55 – 58 show the average time participants spent to answers survey questions.

Table 55: Average time for health domain

Health Domain			
Pair	Table	List	Text
AxB	3:18	3:16	5:26
AxC	3:44	4:31	4:31
AxD	3:51	3:09	5:17
AxE	3:58	3:08	10:30
BxC	3:20	6:03	8:06
BxD	8:27	4:01	8:06
BxE	5:38	4:20	7:33
CxD	3:32	4:00	8:02
CxE	4:15	3:49	7:46
DxE	4:56	3:04	9:44
<b>AVG</b>	<b>4:29</b>	<b>3:56</b>	<b>7:30</b>

Table 56: Average time for financial domain

Financial Domain			
Pair	Table	List	Text
FxG	3:15	4:45	10:55
FxH	6:05	3:59	8:38
FxI	2:05	4:22	6:02
FxJ	2:55	2:57	5:52
GxH	2:21	2:57	3:37
GxI	3:26	2:21	3:29
GxJ	2:54	1:38	3:03
HxI	1:53	4:01	4:12
HxJ	2:35	5:02	5:44
IxJ	3:36	3:33	3:00
<b>AVG</b>	<b>3:06</b>	<b>3:33</b>	<b>5:27</b>

Table 57: Average time for shopping domain

Shopping Domain			
Pair	Table	List	Text
KxL	3:23	4:04	4:02
KxM	3:32	4:01	4:35
KxN	2:59	2:49	5:28
KxO	3:18	3:56	5:31
LxM	3:32	3:15	7:36
LxN	2:51	1:54	3:19
LxO	2:43	2:15	4:02
MxN	3:21	3:00	5:48
MxO	3:27	3:19	5:42
NxO	3:53	2:58	3:39
<b>AVG</b>	<b>3:17</b>	<b>3:09</b>	<b>4:58</b>

Table 58: Average time for networking domain

Social Networking Domain			
Pair	Table	List	Text
PxQ	3:50	2:15	2:46
PxR	2:50	3:13	3:46
PxS	3:13	3:16	5:26
PxT	4:35	3:08	6:03
QxR	2:14	2:21	9:26
QxS	2:13	2:30	2:50
QxT	3:27	1:55	3:27
RxS	2:10	2:08	5:42
RxT	4:32	1:58	4:59
SxT	2:20	2:26	6:11
<b>AVG</b>	<b>3:08</b>	<b>2:31</b>	<b>5:03</b>

The results show that in the highlighted information versions of the survey (table and list), less time was needed by the participant to find answers to a survey's questions compared to the time needed in the original text. We did a t-test analysis between the highlighted information versions of the survey and the original text survey version. In the health domain, we found a significant relationship between the table and text versions with a p-value of 0.0003. Additionally, the relationship between the list and text versions is also significant with a p-value of 0.0001. In the financial domain, our analysis of the results found a significant relationship between the table and text versions with a p-value of 0.006. We also found a significant relationship between the list and text versions with a p-value of 0.009. In the shopping domain, similar results were found for relationships between the table and text versions and the list and text versions respectively, with p-values of 0.001 and 0.0007. In the social networking domain, we found a significant relationship between the table and text versions with a p-value of 0.006. Further, the relationship between the list and text versions is also significant with a p-value of 0.009. Finally, we found that there is no significant difference between the table and list versions as we got p-values of 0.18, 0.2, 0.22, and 0.05 for health, financial, shopping, and social networking domains respectively.

In the next step, we averaged the time of all the domains. The results are summarized in Table 59. The aggregated results show that the participants took less time finding information and answering questions related to two companies when the information is displayed side-by-side in either table or list versions compared to when it is embedded in the original text version of the survey.

Table 59: Average time spent by the participant each domain and each version and the aggregated average

Domain	Table	List	Text
Health	4:19	3:56	7:30
Financial	3:06	3:33	5:27
Shopping	3:17	3:09	4:58
Social Network	3:08	2:31	5:03
<b>AVG</b>	<b>3:27</b>	<b>3:17</b>	<b>5:44</b>

### 5.7.2 Accuracy

Each participant was asked six accuracy test questions to verify his/her ability to find correct information in the different survey versions. These questions were distributed into three survey sections: personal information, controls, and sharing. There were two questions in each section. The first question requires the participant to find which company performs the specific practices of collecting, controlling, or sharing user information. An example of this type of question is, “**Which company collects the user's phone number?**” The other question verifies a participant’s ability to comprehend which of the two companies collects more information, gives more controls and shares information with more parties. An example of this type of questions is, “**Which company gives users more control over the use of their personal information?**” For each pair of companies in each domain, we calculated the average percentage of the correct answers. The following tables show the average accuracy percentages for each pair of companies. Also, the average of all pairs of each version of the survey for each individual domain.

Table 60: The average accuracy of participants getting correct answer about specific information and about the company who collects more data of health domain.

Table			List			Text		
Pair	Specific information	More information	Pair	Specific information	More information	Pair	Specific information	More information
AxB	100%	93%	AxB	100%	93%	AxB	33%	67%
AxC	73%	80%	AxC	87%	93%	AxC	40%	67%
AxD	80%	93%	AxD	89%	100%	AxD	60%	53%
AxE	87%	100%	AxE	100%	87%	AxE	47%	40%
BxC	100%	100%	BxC	93%	80%	BxC	73%	73%
BxD	94%	83%	BxD	89%	89%	BxD	73%	47%
BxE	87%	87%	BxE	80%	93%	BxE	67%	87%
CxD	80%	73%	CxD	93%	87%	CxD	77%	89%
CxE	73%	100%	CxE	60%	80%	CxE	40%	73%
DxE	87%	73%	DxE	67%	80%	DxE	47%	53%
<b>AVG</b>	<b>86%</b>	<b>88%</b>	<b>AVG</b>	<b>86%</b>	<b>88%</b>	<b>AVG</b>	<b>56%</b>	<b>65%</b>

Table 61: The average accuracy of participants getting correct answer about specific information and about the company who collects more data of financial domain.

Table			List			Text		
Pair	Specific information	More information	Pair	Specific information	More information	Pair	Specific information	More information
FxG	100%	87%	FxG	100%	100%	FxG	57%	30%
FxH	80%	80%	FxH	87%	87%	FxH	87%	87%
FxI	100%	100%	FxI	87%	93%	FxI	47%	57%
FxJ	100%	100%	FxJ	80%	100%	FxJ	67%	47%
GxH	80%	100%	GxH	53%	93%	GxH	67%	60%
GxI	80%	93%	GxI	87%	100%	GxI	67%	53%
GxJ	100%	100%	GxJ	87%	100%	GxJ	100%	80%
HxI	100%	100%	HxI	93%	87%	HxI	100%	60%
HxJ	100%	100%	HxJ	80%	87%	HxJ	93%	87%
IxJ	93%	93%	IxJ	93%	100%	IxJ	93%	80%
<b>AVG</b>	<b>93%</b>	<b>95%</b>	<b>AVG</b>	<b>85%</b>	<b>95%</b>	<b>AVG</b>	<b>78%</b>	<b>64%</b>

Table 62: The average accuracy of participants getting correct answer about specific information and about the company who collects more data of shopping domain.

Table			List			Text		
Pair	Specific information	More information	Pair	Specific information	More information	Pair	Specific information	More information
KxL	100%	80%	KxL	87%	67%	KxL	73%	53%
KxM	93%	80%	KxM	87%	87%	KxM	67%	60%
KxN	73%	93%	KxN	60%	93%	KxN	47%	60%
KxO	93%	87%	KxO	87%	80%	KxO	67%	67%
LxM	80%	80%	LxM	87%	87%	LxM	60%	67%
LxN	93%	80%	LxN	87%	87%	LxN	67%	87%
LxO	93%	100%	LxO	80%	67%	LxO	73%	67%
MxN	87%	100%	MxN	73%	93%	MxN	40%	93%
MxO	100%	93%	MxO	73%	80%	MxO	67%	53%
NxO	100%	93%	NxO	87%	87%	NxO	80%	53%
<b>AVG</b>	<b>91%</b>	<b>89%</b>	<b>AVG</b>	<b>81%</b>	<b>83%</b>	<b>AVG</b>	<b>64%</b>	<b>66%</b>

Table 63: The average accuracy of participants getting correct answer about specific information and about the company who collects more data of social networking domain.

Table			List			Text		
Pair	Specific information	More information	Pair	Specific information	More information	Pair	Specific information	More information
PxQ	87%	93%	PxQ	60%	67%	PxQ	27%	53%
PxR	100%	93%	PxR	87%	73%	PxR	47%	60%
PxS	73%	87%	PxS	80%	80%	PxS	87%	80%
PxT	100%	100%	PxT	73%	80%	PxT	60%	53%
QxR	100%	53%	QxR	80%	73%	QxR	53%	67%
QxS	93%	100%	QxS	100%	93%	QxS	73%	67%
QxT	93%	93%	QxT	100%	100%	QxT	53%	40%
RxS	93%	100%	RxS	87%	93%	RxS	40%	53%
RxT	87%	87%	RxT	93%	93%	RxT	40%	67%
SxT	87%	91%	SxT	100%	93%	SxT	60%	53%
<b>AVG</b>	<b>91%</b>	<b>90%</b>	<b>AVG</b>	<b>86%</b>	<b>85%</b>	<b>AVG</b>	<b>54%</b>	<b>59%</b>

Based on the displayed results we can say that participants could find correct information when they are provided with highlighted and side-by-side data compared to finding information when looking to the original text of pair of policies. A t-test was performed to examine the accuracy, and the results show a significant relationship between the table and list versions and the text survey version. However, in most cases there is no significant difference between the two highlighted versions (i.e. the table and the list versions.) Table 64 shows the p-values of the t-test between the versions. For the relationships between the highlighted versions and text version the values were all significant except for that of the relationship between list-text versions in the financial domain. This is due to the fact that the financial domain's original text is displayed as a table showing highlighted information which provided information (to the participants) in a format as easy as the list version.

Table 64: P-value of domains across all survey versions.

Domain	Tested versions	Specific information (p-value)	More information (p-value)
Health	table-text	0.0001	0.003
	table-list	<b>0.46</b>	<b>0.5</b>
	list-text	0.0003	0.001
Financial	table-text	0.01	0.0004
	table-list	<b>0.17</b>	0.0009
	list-text	<b>0.16</b>	0.0009
Shooping	table-text	9.18E-07	0.0005
	table-list	0.001	<b>0.08</b>
	list-text	0.0001	0.001
Social Networking	table-text	0.0002	0.0005
	table-list	<b>0.15</b>	<b>0.15</b>
	list-text	0.0001	0.0007

The overall accuracy percentages of participants' responses to questions across different versions of the survey are displayed in Table 65. The results indicate that the participants were able to find more accurate information when refereeing to the highlighted side-by-side versions than when refereeing to original text version.

Table 65: The accuracy percentage of participants' responses across different versions of the survey.

Table		
Domain	Specific information	More information
Health	86%	88%
Financial	93%	95%
Shopping	91%	89%
Social Network	91%	90%
<b>AVG</b>	<b>90%</b>	<b>90%</b>

List		
Domain	Specific information	More information
Health	86%	88%
Financial	85%	95%
Shopping	81%	83%
Social Network	86%	85%
<b>AVG</b>	<b>85%</b>	<b>88%</b>

Text		
Domain	Specific information	More information
Health	56%	65%
Financial	78%	64%
Shopping	64%	66%
Social Network	54%	59%
<b>AVG</b>	<b>63%</b>	<b>64%</b>

### 5.7.3 Preferences

Based on the results in the previous section stating that participants can more easily accurately find information in the highlighted side-by-side versions than the text version we decided to analyze these versions to determine participant's privacy preferences. We calculated the percentages of participants who prefer the policy that collects less personal information, provides more controls, and shares information with fewer parties. Table 66 shows participants' preferences for both versions across all domains.

*Table 66: Participants' preferences for list and table versions of all domains.*

<b>Domain-version</b>	<b>Prefer less Info. collection</b>	<b>Prefer more Conrtols</b>	<b>Prefer less Sharing</b>
Health- list	90%	100%	77%
Health- table	90%	85%	100%
Financial- list	80%	100%	100%
Financial- table	100%	100%	100%
Shopping- list	70%	100%	85%
Shopping- table	100%	100%	100%
Networking- list	100%	88%	100%
Networking- table	88%	100%	100%
<b>AVG</b>	<b>90%</b>	<b>97%</b>	<b>95%</b>

The data shows that users prefer a company that collects less information, provides more controls over the data, and shares information with fewer parties.

In the next chapter, we will further investigate the data collected from the surveys regarding users' privacy preferences to compare and rank more than two policies.

## CHAPTER 6: Ranking Privacy Policies

### 6.1 Overview

In the previous chapter, we argued that presenting the privacy policies of two companies side-by-side would allow users to compare the policies and make informed decisions. This solution is useful when comparing small numbers of companies, for example, two to four companies. However, this method would be confusing and difficult to use to present the content of five or more companies' privacy policies. In the case of more than five companies, assigning a score to privacy policies is a more appropriate approach in which the user can evaluate the privacy practices of a certain service provider based on the company's assigned score.

Calculating a score can be done by asking users to rate privacy policies by determining how good a given policy is and then using the rating variable along with policy features as training data to build a regression model. The regression model would then be used to predict a score for any new policy. We do not think the user rating approach is an ideal solution to our problem because of the issue of reliability of the rating approach in which assessment of ratings may differ from one user to another. For example, a user may rate every policy with three stars and rate the best with four stars, while another user may rate every policy with one star and rate the best with three stars.

Another possible solution is to use users' preferences to assign scores for privacy policies. Here, when users prefer one policy over the other, their preferences are reliable and

not questionable. In this research, we used “learn to rank”, a machine-learning approach for learning how to derive a ranking function for privacy policies.

Learning to rank is a machine-learning supervised method used for training a model in a ranking task that aims to automatically construct an order from the training data. Learning to rank is employed in several applications such as document retrieval, expert search, definition search, collaborative filtering, question answering, key phrase extraction, document summarization, and machine translation (Li, 2014). In this research, we will implement a learning to rank approach to rank privacy policies by assigning scores to the policies based on users’ preferences.

Recently, in the fields of information retrieval and machine learning, learning to rank has gained the attention of several researchers. Researchers introduced different learning to rank algorithms including AdaRank, Ranking SVM, and RankBoost. Both Ranking SVM and RankBoost are pairwise approaches, however AdaRank is Listwise approach. AdaRank algorithm continually construct a weak ranker and adjust the weight of the training data and finally linearly combines the weak rankers for making ranking prediction. It was mainly introduced to directly optimize any performance measure used in document retrieval. AdaRank is not applicable in problem setting similar to our because we only have one query and no weight adjustment will be needed. Differently, RankBoost and Ranking SVM were designed to optimize loss functions loosely related to the information retrieval performance measures. For example, they train ranking models by minimizing classification errors on instance pairs. Both AdaRank and RankBoost use a process similar to Adaboost to combine a set of weak learners (each with

a weight). Although they can be used to rank policies, the ranking model is quite complicated and is hard for ordinary users to interpret. Ranking SVM only learns one model and is very easy to interpret. Thus, in this dissertation we will apply Ranking SVM in the ranking process (Freund, Iyer, Schapire, and Singer, 2003; Herbrich, Graepel, and Obermayer, 2000; Xu and Li, 2007)

### 6.1.1 Overview of pairwise comparison

We used the previous chapter's data as an example to explain the pairwise comparison approach. We used the data to compare and rank the 20 policies of the training dataset, separating out the five policies from each domain into sets. As the responses to the table format version of the survey reflected the greatest degree of accuracy, we used the table format results in our example.

We used the pairwise comparison method to compare the privacy policies within the same domain. For each pair of policies, the total number of participants' preferences was determined, and then a point was given for the policy with the higher number and a half point for both companies in a tie. Then, we calculated the total points for each policy and defined a score. The policy with the higher score was rated as the most preferable with a value of 1, while the policy with the lowest score was ranked as the least preferable with a value of 5. Tables 67 – 70 present the comparison tables of each of the four domains.

Table 67: Pairwise comparison of the health domain

	A	B	C	D	E	SCORE	RANK
A	-	1	0	0	0	1	4
B	0	-	0	0	0	0	5
C	1	1	-	0.5	0.5	3	2
D	1	1	0.5	-	0	2.5	3
E	1	1	0.5	1	-	3.5	1

Table 68: Pairwise comparison of the financial domain

	F	G	H	I	J	SCORE	RANK
F	-	0	0	1	0	1	3
G	1	-	1	1	1	4	1
H	1	0	-	1	0	2	2
I	0	0	0	-	0	0	4
J	1	0	0	0	-	1	3

Table 69: Pairwise comparison of the shopping domain

	K	L	M	N	O	SCORE	RANK
K	-	1	1	1	1	4	1
L	0	-	0.5	0	0.5	1	4
M	0	0.5	-	0	0	0.5	5
N	0	1	1	-	0.2	2.5	2
O	0	0.5	1	0.5	-	2	3

Table 70: Pairwise comparison of the networking domain

	P	Q	R	S	T	SCORE	RANK
P	-	0.5	0	0	0.5	1	3
Q	0.5	-	0	0	1	1.5	2
R	1	1	-	0.5	1	3.5	1
S	1	1	0.5	-	1	3	1
T	0.5	0	0	0	-	0.5	3

## 6.2 Data collection and labeling

Based on the results of the tabular surveys, we determined policy preferences by referring to the question that asks the participants to choose which company they prefer to use or deal with. As there are five policies in each domain, we had ten pairs per domain and a total of forty pairs for the whole dataset. For each pair, we gathered input of five participants who compared the policies of a pair and specified their preferences. We created a training dataset of policies by computing the differences of the features of the privacy practices for every pair of policies in each domain. We added a label that indicates users' preferences between the two policies, then we added the negative values of the pairs.

In addition to the dataset's features that represent the privacy practices, we added three counting features that represent the amount of data collected, the number of provided controls, and the number of parties with which users' information is shared for each individual policy. These types of information help in results analysis, as explained in section 6.6.

Two different labeling approaches were used to label the pair of policies. The first approach is the binary approach where a pair is labels as "1" if the first policy is preferable, as "-

1” if the second policy is preferable, and “0” if there is no preference between the two policies of a pair. We determined the overall preference of a policy over the other by computing how many participants chose each company. For example, if three participants chose “p1” and only two participants chose “p2”, then “p1” is preferable. Also, if two participants chose “p1”, two participants chose “p2”, and one participant chose “no preference”, then it is determined as “no preference” between the two policies and the pair labeled as “0.” Figure 12 shows an example of binary labeling two policies.

	f1	f2	f3	f4	f5	
p1	1	1	0	1	1	
p2	0	1	1	0	1	
	f1	f2	f3	f4	f5	label
p1-p2	1	0	-1	1	0	1
p2-p1	-1	0	1	-1	0	-1

Figure 13: An example of binary labelling

The other labeling approach, what we called “proportional labeling,” is presented in Figure 14. In this approach, we give one point for a policy chosen as the more preferable option and give a half point for both policies when a participant has no preference. Then we calculate a score for each policy pair by the following formula,  $p_i$  is total point given to a policy.

$$\text{Score}(p_{i-j}) = \frac{p_i}{p_i + p_j}$$

For example, if one participant prefers “p1”, three participants prefer “p2”, and one participant has no preference, then “p1” got 1.5 (1+0.5) points and “p2” got 3.5 (3+ 0.5) points. Then, we label the pair (p1-p2) as 0.3 (1.5/5) and (p2-p1) as 0.7 (3.5/5).

	f1	f2	f3	f4	f5	
p1	1	1	0	1	1	
p2	0	1	1	0	1	
	f1	f2	f3	f4	f5	label
p1-p2	1	0	-1	1	0	0.2
p2-p1	-1	0	1	-1	0	0.7

Figure 14: An example of proportional labelling

### 6.3 Learn to rank

#### Problem definition

Suppose that  $P=\{p_1, p_2, \dots, p_n\}$  is the set of policies for training where  $p_i$  is the  $i$ -th policy. Users provide a set of preferences  $S^*$  in the format of  $p_i \succ p_j$ , where  $\succ$  denotes the order relation, meaning users prefer  $p_i$  to  $p_j$ .

A feature vector  $x_i$  is created from each policy,  $i = 1, 2, \dots, n$ . Let  $x_i = \{x_{i1}, x_{i2}, \dots, x_{in}\}$ . We aim to train a scoring model  $f(x_i)$  that can assign a score to a given policy  $p_i$  based on its feature vector  $x_i$  such that it satisfies preferences in  $S^*$ . That is, if  $p_i \succ p_j$  in  $S^*$  then  $f(x_i) > f(x_j)$ .

We can define the function  $f$  as  $f(x_j) = w \cdot x_j + b$ , where  $\cdot$  is the dot product and

$$f(x_j) = \sum_{i=1}^m w_i \cdot x_j + b, \text{ for policy } p_j$$

Here  $w=(w_1, \dots, w_m)$  is a weight vector. To find the weight vector  $w$ , we transfer the ranking model into pairwise classification. A classifier for ranking orders of policy pairs is created and employed in the ranking of policies. This is represented by the diagram of Figure 15.

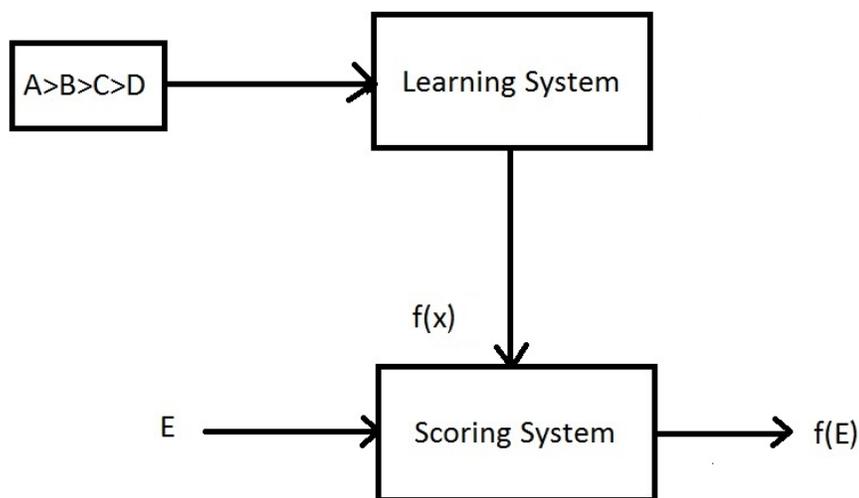


Figure 15: Learning to rank

## 6.4 Feature selection

After the initial pre-processing, the training set consists of 80 instances and 100 features or attributes that represent a company's privacy practices like collecting certain information,

giving control over specific data, or sharing the data with a specific party. In our experiments, we found that applying classification methods to our data does not give reliable results in every case. A major reason for this is the curse of dimensionality, in which we have only 80 instances versus 100 features. To avoid this issue and optimize our prediction results, we applied different feature selection approaches which help to reduce noise, leading to accuracy improvement. Also, these approaches assist in producing interpretable features that results in improving the labeling predictions (Ding & Peng, 2005).

#### 6.4.1 Generalized Feature Selection

One technique we used to reduce the features and get interpretable results is to assign information to a general level. For example, in regards to personal information, if a user's name and birthdate is collected we allocate them to the category of identity information and mark that category with "1" to indicate that the policy collects data in this category. The same method was applied for reducing control and sharing related features. There were four general features for personal information collection: contact, financial, geolocation, and identity. For control related practices, There were five general features: activity, ads and marketing, contact information, communication, and geolocation. For sharing related features, there were three generalization levels: affiliated third parties, non-affiliated third parties, and unspecified third parties.

### 6.4.2 Counting Feature Selection

Another technique we used to reduce dimensionality and improve data analysis is reducing the features to only counting features as we tallied the amount of collected information for the personal information category, the number of provided controls, and the number of parties the information shared.

### 6.4.3 Minimum redundancy feature selection

We used minimum redundancy feature selection approach methods to reduce dimensionality. Minimum redundancy is an approach that was introduced for genomic microarray data to reduce its dimensionality and improve classification predictions, where the number of features is in thousands and much more than the number of instances (Ding & Peng, 2005). The approach is mainly depending on a mutual information  $I$  of two variables as a measure of features relevancy.

$$I(x, y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \quad (1)$$

We applied two microarray feature selection techniques. The first one was MaxRel, which selects the features with the highest relevance to the targeted label or class.

$$\max V_I, \quad V_I = \frac{1}{|S|} \sum_{i \in S} I(h, i) \quad (2)$$

where  $I(h, i)$ , is the mutual information between a feature and targeted class,  $S$  is the number of subset features.

The other one was the minimum redundancy and maximum relevance (mRMR), which minimizes the redundant features and selects the features with highest relevance to the targeted class.

$$\min W_I, \quad W_I = \frac{1}{|S|^2} \sum_{i,j \in S} I(i, j) \quad (3)$$

where  $i$  and  $j$  are two features.

The mRMR feature set is obtained by combining equations (2) and (3), by calculating either the difference or the division between them.

$$\max(V_I - W_I) \quad (4)$$

$$\max(V_I/W_I) \quad (5)$$

In our experiments, we determined the number of relevant features to be 5, 10, and 20 and chose mutual information difference (MID) parameter to determine the feature selection scheme. We used <http://penglab.janelia.org/proj/mRMR/> for the implementation of MaxRel and mRMR feature selection techniques.

## 6.5 Classification methods

The classification problem can be formalized as assigning a score to a pair of policies based on its feature vectors. We have tested the classification methods applying different feature selection techniques described previously to understand which method produces better results. We also tested the classification methods with the different labeling techniques stated

previously. We applied support vector machine and linear/logistic regression classification methods in our experiments, using the Weka Data Mining Software and R for the implementation of the classification.

We initially tested the classification methods using our training dataset, consisting of 80 instances, to get a general idea of how successful our approach would be. We used Weka and R to apply the regression classification with different labeling approaches and feature selection techniques. For the different testing cases, the coefficient of determination  $R^2$  was up to 95%. Thus, we decided to validate our approach by creating a testing dataset, then applying the ranking approach to validate the results.

## 6.6 Ranking Validation

To validate our ranking approach, we collected new data to use as a testing dataset. We randomly chose 12 privacy policies, three from each of the four previously mentioned domains. We extracted information from these policies related to personal information collection, controls, and sharing practices. Then, we created tabular versions of comparison surveys for policies within the same domain in the same way that we did in the previous chapter. There was a total of 12 pairs, three from each domain. We then followed the same survey posting procedures as detailed in the previous chapter as follows. We posted each survey in Amazon Mechanical Turk tool as a HIT. We requested 5 workers for each HIT and paid the workers \$0.75 each. To rank privacy policies, we only considered the general preference data in which

participants were asked which company they preferred based on the general privacy practices of the company.

In the next step, using the surveys' data, we created a data matrix consisting of policy vectors versus privacy practices features including counting features (12x103). Then, we computed the differences between the pair of policies and the negative of the pairs. As a result, the matrix consists of 24 instances (policy pairs) and 103 features. The same labeling approaches and feature selection techniques were applied as those discussed earlier in this chapter.

We applied the two classification methods, SVM and linear/logistic regression, in the process of ranking the policies. We used our original dataset with the 20 policies as the training dataset and the new dataset with the 12 policies as the testing dataset. Tables 71 and 72 show the correlation coefficient values of the classification results.

Table 71:  $R^2$  of SMO regression classification

Feature Selection techniques	Proportional labeling		Binary labeling	
	$R^2$	Degree of Freedom	$R^2$	Degree of Freedom
All	0.67	N/A	0.69	N/A
Counting	0.66	77	0.67	77
Categorical	0.63	65	0.6	65
MaxRel (20)	0.6	60	0.62	60
MaxRel (10)	<b>0.85</b>	<b>70</b>	<b>0.8</b>	<b>70</b>
MaxRel (5)	<b>0.83</b>	<b>75</b>	<b>0.91</b>	<b>75</b>
mRMR (20)	0.79	60	0.72	60
mRMR (10)	0.88	70	0.75	70
mRMR (5)	0.4	75	0.46	75

Table 72:  $R^2$  of linear/logistic regression classification

Feature Selection techniques	Proportional labeling		Binary labeling	
	$R^2$	Degree of Freedom	$R^2$	Degree of Freedom
All	0.35	N/A	0.4	N/A
Counting	0.66	77	0.71	77
Categorical	0.42	65	0.33	65
MaxRel (20)	0.81	60	0.8	60
MaxRel (10)	<b>0.88</b>	<b>70</b>	<b>0.89</b>	<b>70</b>
MaxRel (5)	0.72	<b>75</b>	0.44	<b>75</b>
mRMR (20)	0.73	60	0.64	60
mRMR (10)	<b>0.88</b>	<b>70</b>	<b>0.88</b>	<b>70</b>
mRMR (5)	0.4	75	0.47	75

As the outputs show that the classification models were reliable, we used the predicted output (scores) in the next step to rank the policies. Then we then compare the output ranking with the rankings of using the original labels. We chose the MaxRel (10) output results to apply and validate the ranking approach because it is one of the best two feature selection techniques that had the best output in both classification methods and both labeling approaches. Before the ranking validation, we analyzed the generated classifier model. Figure 16 shows an example of the output model. The results are statistically significant for most of the variable except, health information, ZIP, and SMS.

```

SMOreg
weights (not support vectors):
-      0.5781 * (normalized) billing address
-      0.6003 * (normalized) card number
+      0.1832 * (normalized) health information
-      0.3602 * (normalized) mailing address
-      0.177  * (normalized) SSN
-      0.3602 * (normalized) zip
-      0.1413 * (normalized) Personal information counting
+      0.0837 * (normalized) SMS
+      0.0397 * (normalized) Controls Counting
-      0.6369 * (normalized) Parties Counting
+      1.7737

```

Figure 16: An example of the SMOreg classifier model

The example shows that most personal information-related features have a negative impact on the classifier. That makes sense as users usually do not prefer a company that collects personal information about them. However, the health information feature showed an exception, as it had a positive impact on the classifier. This can be explained by the fact that companies from the health domain collect health information, so it is likely that users are comfortable with those companies collecting that specific personal information because they assume it is required. Another noticeable point is that the model had the number of collected personal information, the number of provided controls, and the number of parties users' information shared with related features as effective features. This means that users care about the total amount of collected information, provided controls, and sharing parties, sometimes regardless of what the specific practice is. Moreover, totals for both personal information and sharing related features showed a negative impact on the classifier, which is an expected result as users usually do not prefer companies that collect a lot of personal information or share their data with a lot parties. Therefore, it is to be expected that companies engage in these practices would receive a lower score. However, features related to the provided controls had a positive impact on the classifier, indicating that users prefer more control over their information.

Tables 73 – 76 show domain specific policies ranking when applying the SMO and liner/logistic regression classifications to a survey's original data to predict the output of proportional and binary labeling.

Table 73: Health domain comparison results

**Pairwise Comparison- Based on Participants' Preferences**

Company	A	B	C	Score	Rank
A	-	1	0.5	1.5	1-2
B	0	-	0	0	3
C	0.5	1	-	1.5	1-2

**Pairwise Comparison- Based on the Prediction of Proportional Labeling**

Company	A	B	C	Score	Rank
A	-	0.91	0.4	1.31	2
B	0.09	-	-0.01	0.08	3
C	0.6	0.9	-	1.5	1

**Pairwise Comparison- Based on the Prediction of Binary Labeling**

Company	A	B	C	Score	Rank
A	-	0.6	-0.14	0.46	2
B	-0.6	-	-0.74	-1.34	3
C	0.15	0.74	-	0.89	1

Table 75: Shopping domain comparison results

**Pairwise Comparison- Based on Participants' Preferences**

Company	K	L	M	Score	Rank
K	-	1	1	2	1
L	0	-	0	0	3
M	0	1	-	1	2

**Pairwise Comparison- Based on the Prediction of Proportional Labeling**

Company	K	L	M	Score	Rank
K	-	0.96	0.96	1.92	1
L	0.03	-	0.5	0.53	3
M	0.04	0.5	-	0.54	2

**Pairwise Comparison- Based on the Prediction of Binary Labeling**

Company	K	L	M	Score	Rank
K	-	0.91	0.95	1.86	1
L	-0.91	-	0.03	-0.88	2
M	-0.95	-0.03	-	-0.98	3

Table 74: Financial domain comparison results

**Pairwise Comparison- Based on Participants' Preferences**

Company	F	G	H	Score	Rank
F	-	0	0	0	3
G	1	-	0	1	2
H	1	1	-	2	1

**Pairwise Comparison- Based on the Prediction of Proportional Labeling**

Company	F	G	H	Score	Rank
F	-	-0.02	-0.02	-0.04	3
G	1	-	0.5	1.5	1-2
H	1	0.5	-	1.5	1-2

**Pairwise Comparison- Based on the Prediction of Binary Labeling**

Company	F	G	H	Score	Rank
F	-	-0.73	-0.92	-1.65	3
G	0.73	-	-0.2	0.53	2
H	0.93	0.2	-	1.13	1

Table 76: Social networking comparison results

**Pairwise Comparison- Based on Participants' Preferences**

Company	P	Q	R	Score	Rank
P	-	1	0.5	1.5	1-2
Q	0	-	0	0	3
R	0.5	1	-	1.5	1-2

**Pairwise Comparison- Based on the Prediction of Proportional Labeling**

Company	P	Q	R	Score	Rank
P	-	0.88	0.35	1.23	2
Q	0.11	-	-0.03	0.08	3
R	0.65	1	-	1.65	1

**Pairwise Comparison- Based on the Prediction of Binary Labeling**

Company	P	Q	R	Score	Rank
P	-	0.4	-0.4	0	2
Q	-0.4	-	-0.8	-1.2	3
R	0.4	0.8	-	1.2	1

These results show that, for proportional labeling, the Spearman’s rank correlation of the ranking approach is 1 where all the policies were ranked correctly. Binary labeling yielded a

0.87 Spearman's rank correlation as two policies, L and M, were ranked incorrectly and reversed.

## 6.7 Discussion and conclusion

With the ranking approach that we proposed, it appears that it is possible to rank companies based on their privacy practices stated in their privacy policies and based on users' privacy preferences. Feature selection techniques helped in improving the interpretability of the results. Our model showed that users mostly prefer companies that collect less information about them, provide them with more control options, and share their information with less parties. Creating versions that simply the information needed to compare companies' practices helps users make better privacy decisions. Users will be knowledgeable and able to avoid privacy risks resulting from information shared with the public or stored in companies' servers, whether intentionally or accidentally.

One suggestion for future research improvement is to increase the dataset scale by including more companies and perhaps more domains. Another suggestion is to consider cross domains, which we found an unnecessary comparison at this point as users typically compare privacy practices among companies providing similar services within each domain, allowing them to choose the company whose practices they prefer.

## CHAPTER 7: Conclusion

Recently, online services have been involved in most of people's daily activities. They use these services to search for specific information, in education, to store their data, and to connect with others. Users usually provide their personal information when it is required to manage their account. Service providers collect and store the provided information in their servers to create users' profiles, to improve their services, or to share it with other parties. Service providers offer privacy policy to describe their privacy practices, usually containing what information they will collect, who will access the collected information, and the purpose of collecting the information.

Studies show that the current privacy policies used by the most of the service providers are difficult to understand by an average user and need a lot of time to read. This leads users to avoid reading the policies and therefore make uninformed privacy decisions, which makes them vulnerable to many privacy risks.

In our thesis, we investigated users' awareness, needs, and concerns regarding their information privacy. We identified the privacy practices that the users are not aware of and the practices that they care about, like information collection, controls and options, and sharing. Then, we analyzed several privacy policies and recognized the text describing related practices and used the analysis results to propose different extraction algorithms. We proposed matching-based approaches to extract personal information and controls related terms. In addition, we proposed a pattern-based approach to extract parties related terms. Using the

extracted information, a side-by-side comparison was proposed where we displayed the information related to privacy practices of pairs of policies in a table. We found that users were able to gain and compare privacy related information easier than when using privacy policy full text. Moreover, a policy ranking approach was proposed to rank policies based on users' preferences and service provider's privacy practices that helps users compare several privacy policies.

Providing the privacy policies comparison and ranking as a tool to users will help them to be informed about their personal information privacy. This can be applied by allowing users to choose two service providers and then display the privacy practices of these providers side-by-side so that a user can decide which one she prefers to use or deal with. Further, allowing users to choose a domain and then display rankings of services providers within the domain in terms of their privacy practices also can help them in deciding which service provider is better than others. Accordingly, helping users to make better privacy decision could prevent them from being vulnerable to privacy breaches.

While policy comparisons and rankings are helpful approaches that improve users' privacy awareness and decisions, they have some limitations. As the approaches were introduced to inform users about the collection of information that can identify them, the approaches do not consider inference information were their identity could be disclosed by combining more than one information which leads to privacy risks. Another limitation is that

the approaches are based on information extraction which is not 100% accurate. Thus, users' preferences and decisions may be effected by the absence or addition of some information.

There are several areas where the work in this dissertation can be extended. **First**, further studies can be done to determine the applicability of the proposed approaches. Users should be questioned about whether they will stop using a service when they know it is not highly ranked. **Second**, this dissertation can be extended by studying factors other than service providers' privacy practices affecting users' decision to use those services, like loyalty to a company, quality of provided services or previous knowledge. **Finally**, providing the users with ranking of privacy policies based on single privacy practice could be a helpful research direction. Some users may be interested in a single privacy practice and need to make a decision based on the ranking of policies of that single practice.

## References

- Aimeur, E., Gambs, S., & Ho, A. (2009, May). UPP: user privacy policy for social networking sites. In *Internet and Web Applications and Services, 2009. ICIW'09. Fourth International Conference on* (pp. 267-272). IEEE.
- Ardagna, C. A., Bussard, L., Di, S. D. C., Neven, G., Paraboschi, S., Pedrini, E., ... & Verdicchio, M. (2009). Primelife policy language.
- Arens, R. J. (2009). Learning to rank documents with support vector machines via active learning.
- Barbaro, M., Zeller, T., & Hansell, S. (2006). A face is exposed for AOL searcher no. 4417749. *New York Times*, 9(2008), 8For.
- Beatty, P., Reay, I., Dick, S., & Miller, J. (2007). P3P adoption on e-Commerce web sites: a survey and analysis. *Internet Computing, IEEE*, 11(2), 65-71.
- Betz, A. (2012). The experiences of adult/child identity theft victims.
- Brodie, C., Karat, C. M., Karat, J., & Feng, J. (2005, July). Usable security and privacy: a case study of developing privacy management tools. In *Proceedings of the 2005 symposium on Usable privacy and security* (pp. 35-43). ACM.

Chaianuchittrakul, C. (2013). *Crowdsourcing Privacy Policy Analysis: Evaluating the Comfort, Readability and Importance of Privacy Policies* (Doctoral dissertation, Carnegie Mellon University).

Consumer Action. (2013). Consumer action "do not track" survey results. Retrieved from [http://www.consumer-action.org/downloads/english/Summary\\_DNT\\_survey.pdf](http://www.consumer-action.org/downloads/english/Summary_DNT_survey.pdf)

Costante, E., den Hartog, J., & Petković, M. (2013). What websites know about you. In *Data Privacy Management and Autonomous Spontaneous Security* (pp. 146-159). Springer Berlin Heidelberg.

Costante, E., Sun, Y., Petković, M., & den Hartog, J. (2012, October). A machine learning solution to assess privacy policy completeness:(short paper). In *Proceedings of the 2012 ACM workshop on Privacy in the electronic society* (pp. 91-96). ACM.

Cranor, L. F., Arjula, M., & Guduru, P. (2002, November). Use of a P3P user agent by early adopters. In *Proceedings of the 2002 ACM workshop on Privacy in the Electronic Society* (pp. 1-10). ACM.

Cullen, T. (2007). *The Wall Street Journal. Complete Identity Theft Guidebook: How to Protect Yourself from the Most Pervasive Crime in America*. Crown Business.

Dinev, T., & Hart, P. (2004). Internet Privacy, Social Awareness, And Internet Technical Literacy. An Exploratory Investigation. *BLED 2004 Proceedings*, 24.

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02), 185-205.

- Donovan, F., & Bernier, K. (2008). *Cyber crime fighters: Tales from the trenches*. Pearson Education.
- Earp, J. B., Anton, A., Aiman-Smith, L., & Stufflebeam, W. H. (2005). Examining Internet privacy policies within the context of user privacy values. *Engineering Management, IEEE Transactions on*, 52(2), 227-237.
- Ehling, Jeff. "Thieves Stealing Barcodes From Pictures Of Event Tickets Posted On Social Media". *ABC13* 2013. Web. 31 Dec. 2015.
- Federal Trade Commission. (2010). *Medical identity theft*. Retrieved January 23, 2011, from <http://www.ftc.gov/bcp/edu/pubs/consumer/idtheft/idt10.shtm>
- Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov), 933-969.
- Gross, R., & Acquisti, A. (2005, November). Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society* (pp. 71-80). ACM.
- Hang, L. I. (2011). A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 94(10), 1854-1862.
- Harris Interactive. (2001). Retrieved from <http://www.harrisinteractive.com>
- Hazari, S., & Brown, C. (2013). An empirical investigation of privacy awareness and concerns on social networking sites. *Journal of Information Privacy and Security*, 9(4), 31-51.

- Herbrich, R., Graepel, T., & Obermayer, K. (1999). Support vector learning for ordinal regression.
- Herbrich, R., Graepel, T., & Obermayer, K. (2000). Large margin rank boundaries for ordinal regression.
- Identity Theft Resource Center. (2007). *Identity theft: The aftermath 2006*. Retrieved November 28, 2007, from [http://www.idtheftcenter.org/artman2/uploads/1/The\\_Aftermath\\_2006\\_2.pdf](http://www.idtheftcenter.org/artman2/uploads/1/The_Aftermath_2006_2.pdf)
- Janda, S., & Fair, L. L. (2004). Exploring consumer concerns related to the internet. *Journal of Internet commerce*, 3(1), 1-21.
- Javelin Strategy & Research. 2006. *2006 Identity Fraud Survey Report*. January.
- Jensen, C., & Potts, C. (2004, April). Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 471-478). ACM.
- Joshi, K. P., Gupta, A., Mittal, S., Pearce, C., Joshi, A., & Finin, T. (2016, December). Semantic approach to automating management of big data privacy policies. In *Big Data (Big Data), 2016 IEEE International Conference on* (pp. 482-491). IEEE.
- Kelley, P. G., Bresee, J., Cranor, L. F., & Reeder, R. W. (2009, July). A nutrition label for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (p. 4). ACM.

- Kelley, P. G., Cesca, L., Bresee, J., & Cranor, L. F. (2010, April). Standardizing privacy notices: an online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems* (pp. 1573-1582). ACM.
- Leon, P. G., Cranor, L. F., McDonald, A. M., & McGuire, R. (2010, October). Token attempt: the misrepresentation of website privacy policies through the misuse of p3p compact policy tokens. In *Proceedings of the 9th annual ACM workshop on Privacy in the electronic society* (pp. 93-104). ACM.
- Leyden, J. (2004). US phishing losses hit \$500 m. *The Register*, 29.
- Li, H. (2014). Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies*, 7(3), 1-121.
- Libert, T. (2015). Privacy implications of health information seeking on the web. *Communications of the ACM*, 58(3), 68-77.
- Lin, J., Amini, S., Hong, J. I., Sadeh, N., Lindqvist, J., & Zhang, J. (2012, September). Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 501-510). ACM.
- Mansfield-Devine, S. (2015). The Ashley Madison affair. *Network Security*, 2015(9), 8-16.
- McDonald, A. M., & Cranor, L. F. (2008). Cost of reading privacy policies, the. *ISJLP*, 4, 543.

- McDonald, A. M., Reeder, R. W., Kelley, P. G., & Cranor, L. F. (2009, January). A comparative study of online privacy policies and formats. In *Privacy enhancing technologies* (pp. 37-55). Springer Berlin Heidelberg.
- Mont, M. C. (2004). Dealing with privacy obligations in enterprises. In *ISSE 2004—Securing Electronic Business Processes* (pp. 198-208). Vieweg+ Teubner Verlag.
- Mont, M. C., Thyne, R., & Bramhall, P. (2005). *Privacy enforcement with HP Select Access for regulatory compliance*. Technical Report HPL-2005-10, HP Laboratories Bristol, Bristol, UK.
- Moores, T. (2005). Do consumers understand the role of privacy seals in e-commerce?. *Communications of the ACM*, 48(3), 86-91.
- Nowak, G. J., & Phelps, J. (1992). Understanding privacy concerns. An assessment of consumers' information-related knowledge and beliefs. *Journal of Direct Marketing*, 6(4), 28-39.
- Pollach, I. (2007). What's wrong with online privacy policies?. *Communications of the ACM*, 50(9), 103-108.
- Reeder, R. W., Kelley, P. G., McDonald, A. M., & Cranor, L. F. (2008, October). A user study of the expandable grid applied to P3P privacy policy visualization. In *Proceedings of the 7th ACM workshop on Privacy in the electronic society* (pp. 45-54). ACM.
- Reisman, Suzanne. "The Murder Of Asia McGowan". *BlogHer* 2009. Web. 31 Dec. 2015.

Schmidt, S. W., & McCoy, M. R. (2008). *The Silent Crime: What You Need to Know about Identity Theft*. Twin Lakes Press.

Singh, R. I., Sumeeth, M., & Miller, J. (2011). A user-centric evaluation of the readability of privacy policies in popular web sites. *Information Systems Frontiers*, 13(4), 501-514.

Smith, H. J., Dinev, T., & Xu, H. (2011). Information privacy research: an interdisciplinary review. *MIS quarterly*, 35(4), 989-1016.

T. Reilly. Press release: AG warns consumers: Do not give your private financial information to anyone calling or emailing pretending to be investigating the Boston Globe security breach. URL <http://www.ago.state.ma.us/sp.cfm?pageid=986&id=1602>, 2006.

The Center for Information Policy Leadership. Multi-Layered Notices Explained, 2004.

[http://www.hunton.com/files/tbls47Details/FileUpload265/1303/CIPLAPEC\\_Notices\\_White\\_Paper.pdf](http://www.hunton.com/files/tbls47Details/FileUpload265/1303/CIPLAPEC_Notices_White_Paper.pdf).

Tsai, J. Y., Egelman, S., Cranor, L., & Acquisti, A. (2011). The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research*, 22(2), 254-268.

Tufekci, Z. (2008). Can you see me now? Audience and disclosure regulation in online social network sites. *Bulletin of Science, Technology & Society*, 28(1), 20-36.

Walters, W., & Betz, A. (2012). Medical identity theft. *Journal of Consumer Education*, 75.

Westin, A. (2001). *Opinion surveys: What consumers have to say about information privacy*.

Prepared Witness Testimony, The House Committee on Energy and Commerce.

Xu, J., & Li, H. (2007, July). Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 391-398). ACM.

Yao, M. Z., Rice, R. E., & Wallis, K. (2007). Predicting user concerns about online privacy. *Journal of the American Society for Information Science and Technology*, 58(5), 710-722.

Young, A. L., & Quan-Haase, A. (2009, June). Information revelation and internet privacy concerns on social network sites: a case study of facebook. In *Proceedings of the fourth international conference on Communities and technologies* (pp. 265-274). ACM.

Yu, W. D., & Murthy, S. (2007, July). PPMLP: A special modeling language processor for Privacy policies. In *Computers and Communications, 2007. ISCC 2007. 12th IEEE Symposium on* (pp. 851-858). IEEE.

Zimmeck, S., & Bellovin, S. M. (2014, August). Privee: An architecture for automatically analyzing web privacy policies. In *Proceedings of the 23rd USENIX Security Symposium*.

