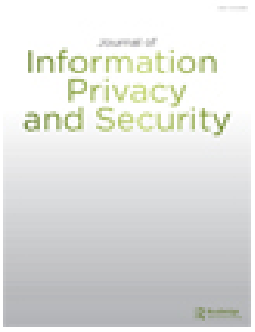

Routledge
Taylor & Francis Group

Journal of Information Privacy and Security

ISSN: 1553-6548 (Print) 2333-696X (Online) Journal homepage: http://www.tandfonline.com/loi/uips20

# A study of web privacy policies across industries

Razieh Nokhbeh Zaeem & K. Suzanne Barber

**To cite this article:** Razieh Nokhbeh Zaeem & K. Suzanne Barber (2017): A study of web privacy policies across industries, Journal of Information Privacy and Security, DOI: 10.1080/15536548.2017.1394064

**To link to this article:** https://doi.org/10.1080/15536548.2017.1394064

Published online: 13 Nov 2017.

JOURNAL OF INFORMATION PRIVACY AND SECURITY
https://doi.org/10.1080/15536548.2017.1394064

# A study of web privacy policies across industries

Razieh Nokhbeh Zaeem and K. Suzanne Barber

Center for Identity, University of Texas at Austin, Austin, USA

**ABSTRACT**
Today, more than ever, companies collect their customers' Personally Identifiable Information (PII) over the Internet. The alarming rate of PII misuse drives the need for improving companies' privacy practices. We thoroughly study privacy policies of 600 companies (10% of all listings on NYSE, Nasdaq, and AMEX stock markets) across industries and investigate 10 different privacy pertinent factors in them. The study reveals interesting trends: for example, more than 30% of the companies still lack privacy policies, and the rest tend to collect users' information but claim to use it only for the intended purpose. Furthermore, almost one out of every two companies provides the collected information to law enforcement without asking for a warrant or subpoena. We found that the majority of the companies do not collect children's PII, one out of every three companies lets users correct their PII but does not allow complete deletion, and the majority post new policies online and expect the user to check the privacy policy frequently. The findings of this study can help companies improve their privacy policies, enable lawmakers to create better regulations and evaluate their effectiveness, and finally educate users with respect to the current state of privacy practices in an industry.

**KEYWORDS**
Privacy; policy; study; industries

## Introduction

The ever-growing use of the Internet and the collection of Personally Identifiable Information (PII) over it has raised concerns for close to two decades (Culnan, 1999; FTC, 1998). In particular, the problem of how companies handle users' PII collected over the Internet involves three main players: companies, regulators, and users.

*Companies*, across industries, currently are faced with tough decisions when constructing their privacy policies. On the one hand, many business models are built on collecting, using, sharing, and selling personal information. Such information can be profitable for the company and can be leveraged to improve their product offerings and consumer-facing services. On the other hand, collecting and storing personal information about consumers carries considerable risk, as evidenced by the financial and public relations fallout from high-profile hacks experienced by Target and Sony. Data breaches are occurring at alarming rates, and the public relations and financial fallout from such hacks can be massive. Companies must assess the balance of these risk/value propositions as they construct their privacy policies.

In response to high profile data breaches, *regulators* and policy makers—the second important player—have employed two lines of strategy: (1) holding corporations liable for breaches, imposing fines, and sanctions on organizations that handled consumer data inappropriately and (2) attempting to increase the transparency of privacy and data management practices in privacy policies. The regulators, however, must constantly assess the current state of privacy policies across industries and evaluate the effects of the regulations they establish (Romanosky, Telang, & Acquisti, 2011).

**CONTACT** Razieh Nokhbeh Zaeem razieh@identity.utexas.edu; K. Suzanne Barber sbarber@identity.utexas.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/UIPS.

Finally, in the face of companies' carefully constructed privacy policies and regulators' endeavors to encourage transparency in privacy policies, *users* have neither the time (Kohavi, 2001; McDonald & Cranor, 2008; Meinert, Peterson, Criswell, & Crossland, 2006; Milne & Culnan, 2004) nor the inclination (Graber, Alessandro, & Johnson-West, 2002; Milne, Culnan, & Greene, 2006) to read privacy policies thoroughly, choosing instead to agree absentmindedly to the various privacy policies. More than ever, consumers need information to help them compare what a privacy policy offers with the status quo (e.g., average privacy practices among the companies that provide similar services).

This study seeks to help all of the above groups—companies, regulators, and users—by offering an extensive and in-depth investigation of privacy policies across industries. The study considers 600 companies (10% of all the companies listed on the New York Stock Exchange (NYSE), Nasdaq, and American Stock Exchange (AMEX) stock markets) across more than ten industries and examines ten major privacy pertinent factors in their privacy policies.

As the next section explains, many researchers have examined privacy policies in the past two decades. However, they almost unanimously focus on the same set of four Fair Information Practice Principles (FIPPs): Notice, Choice, Access, and Security. In addition, there are few comprehensive studies across industries. This work seeks to fill the void by comprehensively investigating a large set of privacy policies, randomly selected across industries, and extending the privacy pertinent factors it considers beyond the standard set of four FIPPs.

## Related work

Research suggests that users are illiterate with regards to privacy policies. For example, a survey (Turow & Center, 2003) showed that 57% of U.S. adults wrongly believed that when a website has a privacy policy, it would not share user data. As another example, self-report studies reveal that less than half of the users surveyed have *ever* read a privacy policy (Meinert et al., 2006) and only 4.5% claim to always read them (Milne & Culnan, 2004). Even worse, the more reliable server side observation of websites shows that only 1% or less of users click on a website's privacy policy (Kohavi, 2001). This level of illiteracy leaves users vulnerable to the misuse of their PII. More recent studies found that informing users, for example by clearly displaying privacy policies (Pan & Zinkhan, 2006; Tsai, Egelman, Cranor, & Acquisti, 2011), motivates them to incorporate privacy into their online purchase decisions. Therefore, it is vital to educate users through tools, information, and statistics about the status quo of privacy policies.

We believe this work is the first of its kind to study privacy policies in-depth across industries. In this section, we first cover closely related work and then discuss theoretical work performed on privacy concerns and policies.

### *Similar studies*

The Federal Trade Commission (FTC) has provided several reports on online privacy practices since 1995. Its 1998 report (FTC, 1998) on U.S. commercial websites' privacy disclosures revealed that while 92% of websites were collecting PII, only 14% disclosed any privacy policies. In its 2000 report (FTC, 2000), the FTC investigated a group of 335 websites chosen randomly and another group of 100 most busiest websites, both groups from the U.S. market. The FTC noted that a vast majority of the websites studied collected some PII, e.g., 97% of the random sample and 99% of the busiest websites asked for email addresses. The same study found that, in year 2000, 88% of the random group and all of the 100 busiest websites disclosed some form of their privacy policy. In the same time frame, a survey of 100 most heavily used websites (Culnan, 1999) focused on Notice, Choice, Access, and Security—the four FIPPs. These studies date back 15 years. In addition, they differ from ours in several ways: they did not partition privacy policies based on industries, nor did they try to cover various industries. They did not include as many privacy policies as we did, and finally, they concentrated only on the four FIPPs.

Since the first round of studies performed by the FTC, many researchers have analyzed the content of privacy policies in various ways. Many have investigated privacy policies with respect to a set of factors. The vast majority of such investigations, however, has considered only the FTC's four FIPPs. These investigations have considered different sample sets of privacy policies; a study of 32 Dow Jones Corporations (Li, Stweart, Zhu, & Ni, 2012), the investigation of Fortune e- 50 (Ryker, Lafleur, McManis, & Cox, 2002) companies, a cross-cultural analysis of 150 companies selected from Forbes' Global 2000 company list (Zhang, Toru, & Kennedy, 2007), an examination of 183 companies with headquarters in the Middle East (Shalhoub, 2006), and a study using 600 health websites (Rains & Bosch, 2009) all concentrated on the four FIPPs. Occasionally, some have taken into account a broader range of factors. For instance, Cha (2011) employed the guidelines of the EU Data Protection Directive, requiring all seven privacy components: Notice, Choice, Onward Transfer, Access, Security, Accountability, and Data Integrity. While considering the FIPPs and the EU Data Protection Directive has shed some light on the current state of privacy policies, a wider range of factors, like what we considered in this work, would allow a deeper analysis of the policies. In fact, previous work suggests (Rains & Bosch, 2009) a study like ours that considers difference privacy characteristics (e.g., email and cookies) be undertaken.

In the study that resembles our work the most, Liu and Arnett (2002) examined websites of the Global 500 and showed that only 61% of companies in the United States had posted privacy policies. They extended their search effort for companies' privacy policies by e-mailing the companies and asking for their policies, when one could not be found online. Even with that extra effort, they showed that only 24% of the websites without posted privacy policy that they contacted indeed did have a policy elsewhere. Similar to our work, they also considered different market sectors according to Fortune market sectors. They indicated that the entertainment, health care, soaps and cosmetics, and computer sectors have a large percentage (more than 80%) of websites that have privacy policies. On the other hand, companies in publishing and printing and shipping have no privacy policy. The major shortcoming of this work compared to ours, apart from being published in 2002, is that they too used only the four FIPPs.

A more recent trend of studies investigates privacy policies too. An investigation of the Platform for Privacy Preferences (P3P) privacy policies, for example, considered 5,000 websites and their XML-based machine-readable P3P-formatted policies (Cranor, Egelman, Sheng, McDonald, & Chowdhury, 2008). Most recently, an evaluation of the U.S. financial institutions' privacy notices (Cranor, Leon, & Ur, 2016) automatically evaluated well-formatted privacy policies of more than 6,000 financial institutions.

The financial institutions' privacy policies follow a very well-defined format that can be automatically interpreted. Both of these studies automatically reviewed very well-formatted privacy policies. Our work, however, studied natural-language free-format privacy policies, usually pages long. The in-depth analysis of such general privacy policies could not easily and accurately be automated.

Among the previous work in which humans read policies, the biggest corpus of this level of comprehensiveness we could find available online includes 115 privacy policies (Usable Privacy, 2016; Wilson et al., 2016). Furthermore, researchers, e.g. Bhatia, Breaux, & Schaub (2016), have utilized data mining techniques on privacy policies, which makes the task of investigating more policies easier but less accurate.

Finally, note that, in this study, we selected companies from the U.S. market. Therefore the results should be interpreted in the appropriate context. For example, the European tradition views personal privacy as a "human" right, as opposed to how it is viewed as a "consumer" right in the United States (Brown & Layne Blevins, 2002). A separate study of European companies should be performed to judge the level of comprehensiveness and accuracy of privacy policies in the Europe.

## Theoretical work on privacy concerns and policies

It should be noted that there is a body of theoretical work that models and evaluates privacy concerns. A widely cited work of Malhotra, Kim, and Agarwal (2004) uses social contract theory to offer a framework on the dimensionality of Internet users' privacy concerns. Recent work of Zahir

Irani, Sipior, Ward, and Connolly (2013) looks at Malhotra's work and accesses the continued applicability of it and concludes that the Malhotra's work is not the valid scale to employ in measuring information privacy concerns.

Another study (Storey, Kane, & Schwaig, 2009) examines the privacy policies of the Fortune 500 companies to assess the substance and quality of their stated information practices. In that work, six factors are identified that indicate the extent to which a firm is dependent upon consumer personal information, and therefore more likely to develop high-quality privacy statements. While the theoretical work on privacy concerns (Malhotra et al., 2004; Zahir Irani et al., 2013) is not directly related to our study of privacy policies, we plan to extend our work by inspecting the relationship between the comprehensiveness of privacy policies and the company's consumer information dependency (Storey et al., 2009) as part of the future work.

## Company selection

We now turn our attention to the methodology of selecting companies, finding their privacy policies, and investigating them. A scientific methodology of selecting companies is central to conducting a comprehensive and generalizable study of their privacy policies. We aimed for a selection of companies that:

1. Were reputable companies, i.e., listed by well-known stock exchanges.
2. Were categorized based on a standard and commonly used industrial classification.
3. Evenly covered a wide range of categories and industries across that classification.

To achieve a selection that meets these goals, we focused on the companies listed by NYSE, Nasdaq, and AMEX stock markets using Industry Classification Benchmark (ICB).

### *NYSE, Nasdaq, and AMEX*

The NYSE, Nasdaq, and the AMEX[1] are American stock exchanges and are respectively the first, second, and third largest stock exchanges by market capitalization in the Unites States. The Nasdaq Company List includes companies listed on Nasdaq, as well as NYSE and AMEX. As of the date of this article, the companies listed by these three stock markets add up to 6,500 companies worldwide, most of them (5,717 companies) in North America Nasdaq.

We used this company list for selecting companies to study. Note that these companies are publicly traded. The company list uses Industry Classification Benchmark and includes the name, the ICB industry of the company, and the link to its websites.

### *Industry Classification Benchmark*

ICB is an industry classification taxonomy. It segregates markets into 10 industries which are in turn partitioned into 19 super-sectors, further divided into 41 sectors and finally into 114 sub-sectors. Each company is allocated to a sub-sector that most closely resembles its majority source of revenue, and consequently to the corresponding sector, super-sector, and industry (ICB, 2006). Over 70,000 companies and 75,000 securities worldwide are categorized by ICB. ICB is used globally, including by NYSE, Nasdaq, and AMEX (ICB, 2006).

The version of ICB that Nasdaq uses for its company listing, however, is slightly different than the original ICB (ICB, 2006; Nasdaq, 2015) ICB: it adds a new industry (Transportation), and also lists two additional industries (Miscellaneous and N/A) for companies that do not fit solely under one of the other industries. We use the Nasdaq version (with 13 industries) in this study and hereafter refer to it simply as ICB.

[1] AMEX was acquired by NYSE in 2008. It has become NYSE AMEX since 2009.

## Study

In this study, we evaluated the privacy policies of 600 companies. The NYSE, Nasdaq, and AMEX company list contains a total of 5,717 companies in North America (the United States, Canada, and Mexico) as of the date of this article. We selected 600 companies from this list, covering slightly more than 10%.

## Company selection method

Table 1 shows the total number of North American companies listed by NYSE, Nasdaq, and AMEX under each industry. For this study, we randomly selected 10% of each industry using an in-house random number generator. However, some industries had fewer companies listed (e.g., only 72 companies under Transportation). In such sparse industries, selecting only 10% (for example, only seven Transportation companies) would yield very small sample sets. Thus, for each industry, we selected either 10% of the companies or 20 companies, whichever was greater.

## Finding online privacy policies

Once we had the list of companies to study, we proceeded to find the URLs of their privacy policies. We reached the company's website using the link posted on the NYSE, Nasdaq, and AMEX company list. If that link was broken, we performed a Google search with the company name, manually locating the company's website. To get to the privacy policy, we searched for the word "Privacy" on the English version of the company's homepage. If we could not find the privacy policy in this way we performed a Google search for "Privacy" only on the company's website (using the "site" advanced option of the search query). We then manually located the correct URL of the company's online privacy policy. It is important to note that some companies maintain a general privacy policy that governs the overall collection and use of information by the company as well as an online privacy policy that focuses on the collection and use of PII over the Internet. When both of these types of privacy policies were found online, we selected the online privacy policy (and not the general privacy policy) in the manual selection.

## Privacy pertinent factors studied

A privacy policy is usually a lengthy and technical document. We need a list of the most important factors to study in a policy. In order to choose the factors for this study, we evaluated related work and performed a survey.

**Table 1.** Total number of companies and number of companies studied in each industry.

| Industry | Total | Studied |
|---|---|---|
| Finance | 870 | 87 |
| Consumer services | 747 | 75 |
| Technology | 511 | 51 |
| Capital goods | 356 | 36 |
| Basic industries | 299 | 30 |
| Transportation | 72 | 20 |
| Consumer Non-durables | 199 | 20 |
| Consumer durables | 132 | 20 |
| Healthcare | 669 | 67 |
| Public utilities | 246 | 25 |
| Energy | 296 | 30 |
| Miscellaneous | 133 | 20 |
| N/A | 1187 | 119 |
| SUM | 5717 | 600 |

### Previous work on privacy factors

The Organization for Economic Co-operation and Development (OECD) is one of the first to provide Guidelines on the Protection of Privacy, including eight privacy principles (Regard, 1980): Collection Limitation Principle, Data Quality Principle, Purpose Specification Principle, Use Limitation Principle, Security Safeguards Principle, Openness Principle, Individual Participation Principle, and Accountability Principle.

The FTC recommends that privacy policies follow FIPPs (FTC, 2000): Notice, Choice, Access, Security, and Enforcement. We also reviewed public submissions and staff reports from several workshops and roundtables that the FTC held in 2010 and 2012. In these workshops, professors of law, public policy, and computer science, along with representatives of the FTC and the Electronic Frontier Foundation (EFF), concentrated on what users *should* want to know about how companies handle their PII (FTC, 2010, 2012). The workshops suggested these privacy factors: Aggregation, Encryption, Third Party Sharing, Sharing with Law Enforcement, Security, Access, Control, Usage, Ads, Retention, and Location.

More recent work (Usable Privacy) (Wilson et al., 2016) defines these categories for annotating privacy policies: First Party Collection/Use, Third Party Sharing/Collection, User Choice/Control, User Access/Edit/Deletion, Data Retention, Data Security, Policy Change, Do Not Track, and finally International/Specific Audiences.

We also looked up several online services like Disconnect Me Privacy Icons (Disconnect Me., 2014) which includes: Expected Use, Expected Collection, Precise Location, Data Retention, Do Not Track, Children Privacy, SSL Support, Heartbleed, and TRUSTe Certification.

### Factors survey

Since it was not practical to include all the privacy factors we gathered from the literature, we surveyed privacy experts to identify factors that are most important when summarizing a privacy policy. We surveyed 16 full time employees and graduate students of the Center for Identity at UT Austin, who actively work in the field of privacy and security. The participants were asked to score each of the potential factors from 1 to 4. The full questionnaire is shown in the appendix. Using the results of the survey, we selected the factors that this study evaluates and the questions it answers about them.

### List of privacy pertinent factors

We enlisted ten privacy questions to answer about each privacy policy.

(1) How does the site handle your email address?
(2) How does the site handle your credit card number and home address?
(3) How does the site handle your Social Security number?
(4) Does the site use or share your Personally Identifiable Information for marketing purposes?
(5) Does the site track or share your location?
(6) Does the site collect Personally Identifiable Information from children under 13?
(7) Does the site share your information with law enforcement?
(8) Does the site notify you or allow you to opt out when their privacy policy changes?
(9) Does the site allow you to edit or delete your information from its records?
(10) Does the site collect or share aggregated data related to your identity or behavior?

The answers to these questions are mapped to three levels of risk: red (high risk), yellow (medium risk), and green (low risk). Table 2 (from (Nokhbeh Zaeem, German, & Barber, 2015)) shows the

Table 2. Risk levels for privacy pertinent factors.

| Factor | Green Risk Level | Yellow Risk Level | Red Risk Level |
|---|---|---|---|
| (1) Email Address | Not asked for | Used for the intended service | Shared w/third parties |
| (2) Credit Card Number | Not asked for | Used for the intended service | Shared w/third parties |
| (3) Social Security Number | Not asked for | Used for the intended service | Shared w/third parties |
| (4) Ads and Marketing | PII not used for marketing | PII used for marketing | PII shared for marketing |
| (5) Location | Not tracked | Used for the intended service | Shared w/third parties |
| (6) Collecting PII of Children | Not collected | Not mentioned | Collected |
| (7) Sharing w/Law Enforcement | PII not recorded | Legal docs required | Legal docs not required |
| (8) Policy Change Notice | Posted w/opt out option | Posted w/o opt out option | Not posted |
| (9) Choice (Control) of Data | Edit/delete | Edit only | No edit/delete |
| (10) Data Aggregation | Not aggregated | Aggregated w/o PII | Aggregated w/PII |

risk levels for each of the privacy pertinent factors. If a policy skips a privacy pertinent factor altogether, the red level is assigned to it for that factor.

A team of seven privacy experts, graduate, and undergraduate students read each of the policies, totaling close to one million words, and scored each policy according to Table 2 using the red/yellow/green levels. We performed quality control by assigning every policy to two team members and comparing and resolving disagreements between the team members for the first 15% of privacy policies. It is important to note that the ground truth of how a company deals with users' PII is assumed to be what its privacy policy states. Matching the practice of the company with its privacy policy is beyond the scope of this article.

## Findings

The first astonishing finding of this study is the percentage of the websites that completely lack a Web privacy policy. Across all industries, anywhere between 20% and 50% of the companies do not even have an online privacy policy on their website. The Energy industry is particularly lacking in this respect. Other studies, mentioned in the related work, report similar numbers on the percentage of websites without a privacy policy. Figure 1 shows the percentage of the companies we considered in three groups:

(1) Companies with *No Website*
(2) Companies with *No Privacy Policy* on their website
(3) Companies with a privacy policy, i.e., *Policy Studied*

In the rest of this section, we focus on the third group. When comparing industries, we usually skip making statements about the Miscellaneous and N/A industries.

### *Email address*

Figure 2 shows the distribution of risk levels across industries when answering the question regarding handling e-mail addresses. The majority of companies in any industry (81% in total, ranging from 53% to 93% across industries) asks for users' e-mail addresses but claims to use it only for the intended service. The Basic Industries asks for e-mail addresses least frequently. Consumer Durables is the riskiest industry when it comes to selling/sharing e-mail addresses.

### *Credit card number*

As Figure 3 depicts, many companies are at the green risk level with respect to the credit card and other billing information, i.e., they do not ask for such information online because they do not need it. The ones

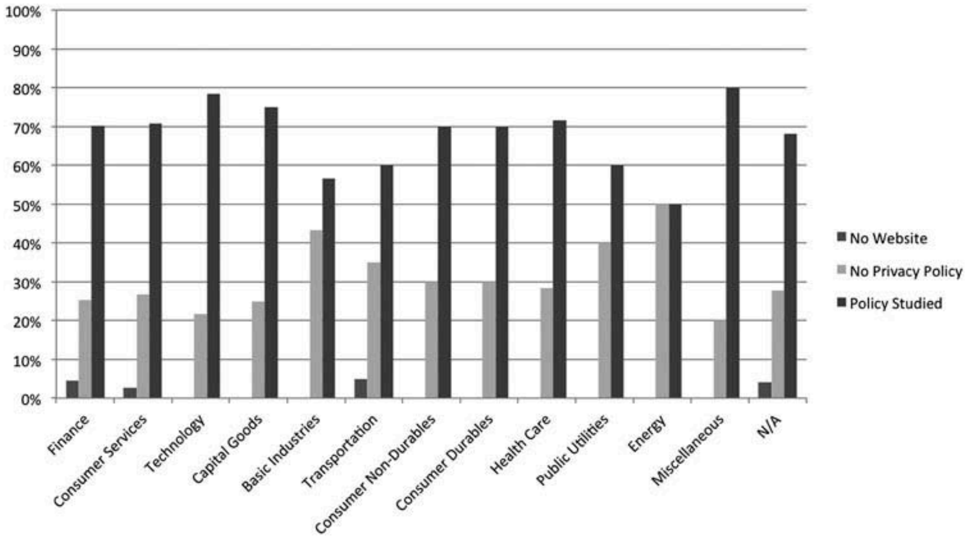

Figure 1. Percentage of companies that have privacy policies.

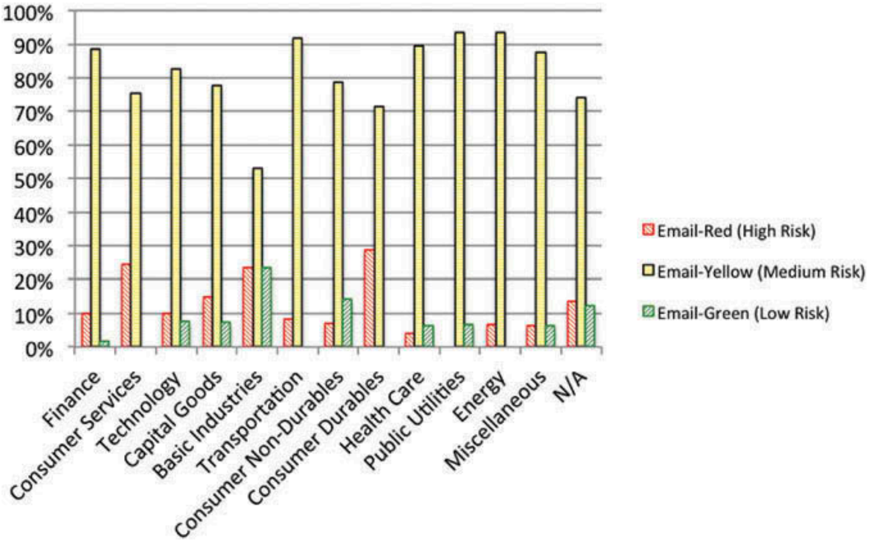

Figure 2. Risk level: Email address.

that do ask for billing information, however, claim to use it for the intended service only. Nonetheless, high profile breaches raise a red flag when it comes to the companies that collect billing information.

## Social security number

Figure 4 shows that many companies do not require users to provide their Social Security Numbers. Yet, there exist companies that do ask for this extremely valuable piece of information. Most noticeably, financial companies (e.g., banks) commonly ask for Social Security Numbers. It is important to note that these companies perform tasks (e.g., tax reporting) that do necessitate the collection of Social Security Numbers.

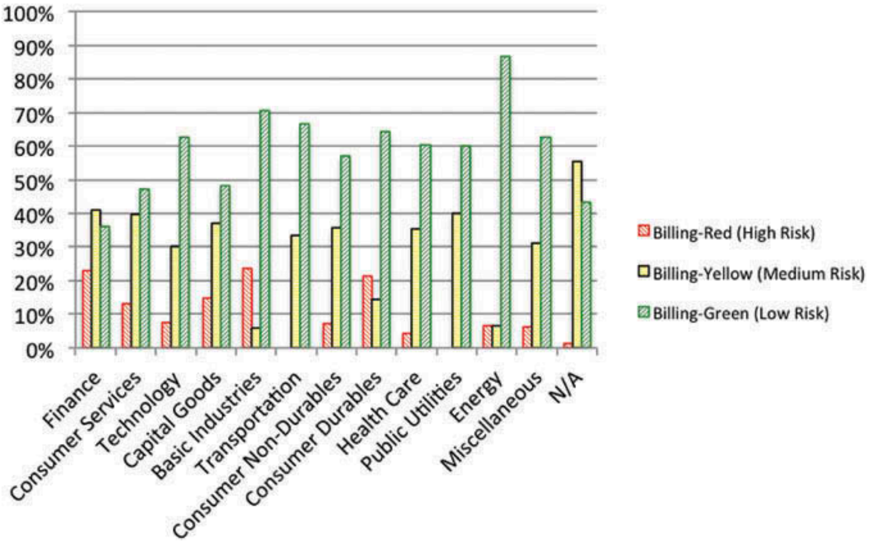

**Figure 3.** Risk level: Credit card number.

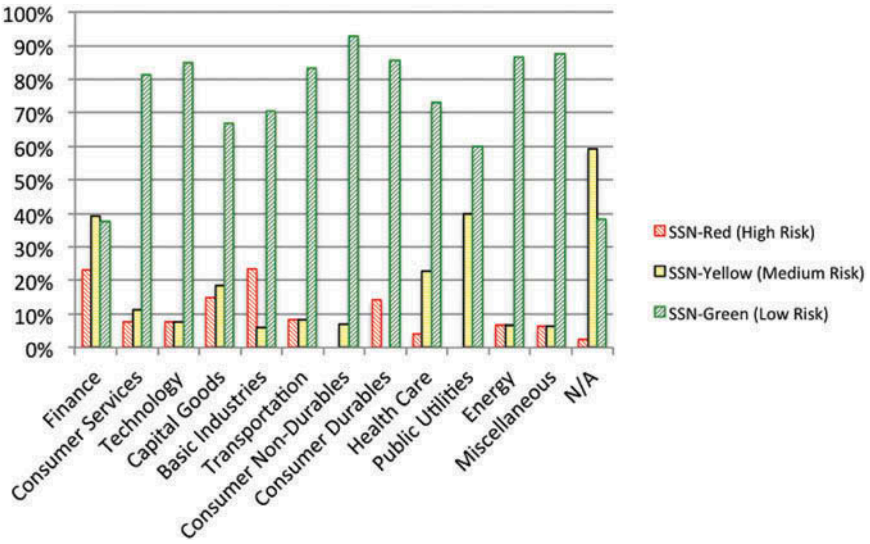

**Figure 4.** Risk level: Social security number.

## *Ads and marketing*

As Figure 5 shows, across all industries, companies use users' PII to serve ads, at the very least to promote their own products. In total, 64% utilize the collected PII to advertise their own products and services and 19% share PII with advertisers on top of that to promote other products and services too. Energy companies use PII to serve ads less than others, and thus have the highest percentage of policies with the green risk level for this privacy pertinent factor. The Finance and Consumer Non-Durables sectors use PII to serve ads most, with more than 70% of their companies having the yellow risk level. Companies in the Consumer Services industry were found to sell PII for marketing and advertisement more than others.

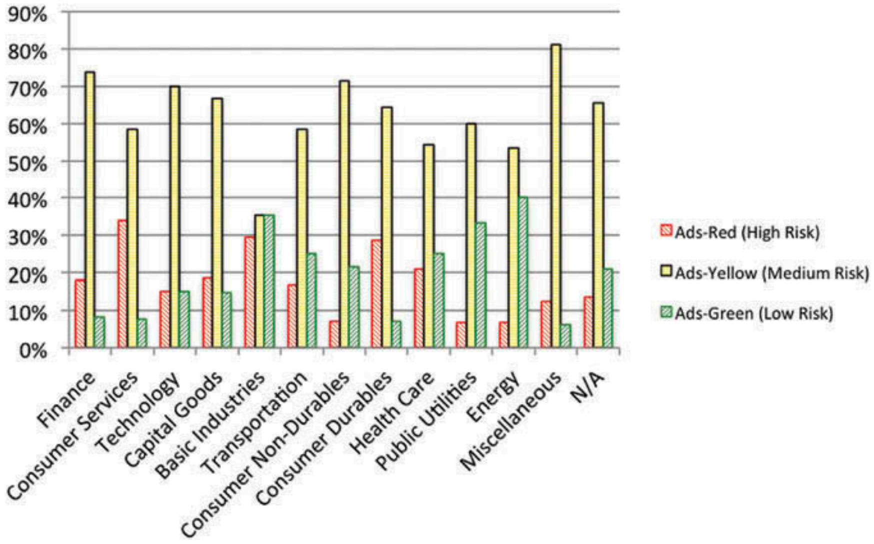

**Figure 5.** Risk level: Using PII for ads and marketing.

## Location

As Figure 6 demonstrates, Consumer Services companies track exact GPS location most often, albeit for the intended service.

## Collecting PII of children

We found that, because of the existing regulations in the United States (COPPA, 1998), companies are vigilant when it comes to collecting Children's PII: only 13% of the total collect the information of children under 13 (see Figure 7).

## Sharing with law enforcement

Another very interesting finding of this study is shown in Figure 8. Almost all companies collect some PII that could be used by law enforcement, and many (45% of the total) would share it with law enforcement without asking for a warrant/subpoena.

## Policy change notice

We found that companies commonly (63% in total) are at the yellow risk level for this factor, i.e., they only post new policies online and continuing to use the website indicates users' implicit agreement (Figure 9).

## Choice (control) of data

As seen in Figure 10, many companies let the users edit their information. Surprisingly, 35% allow editing but do not let the users entirely delete their records. It is important to note that among these are companies that allow the user to delete his or her record, but claim that the record might sill exist in archives and is impossible to fully delete because of technical difficulties.

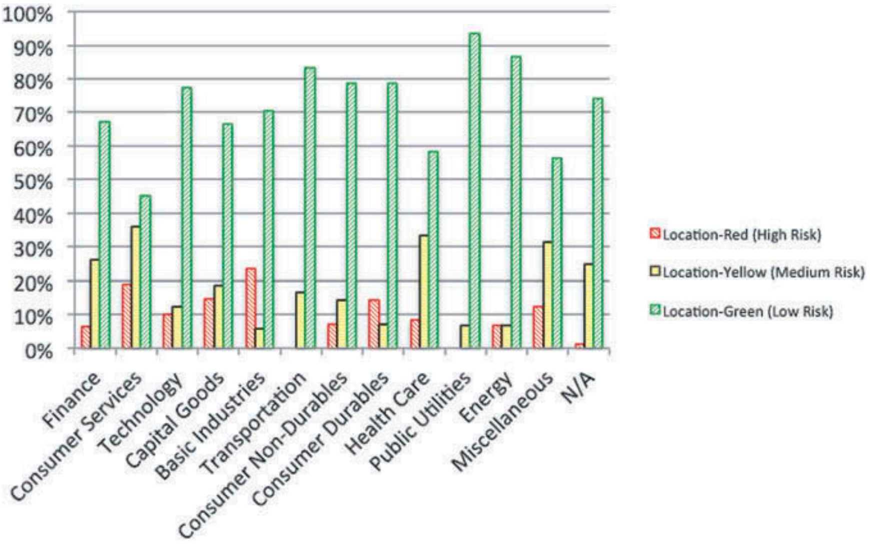

**Figure 6.** Risk level: Location.

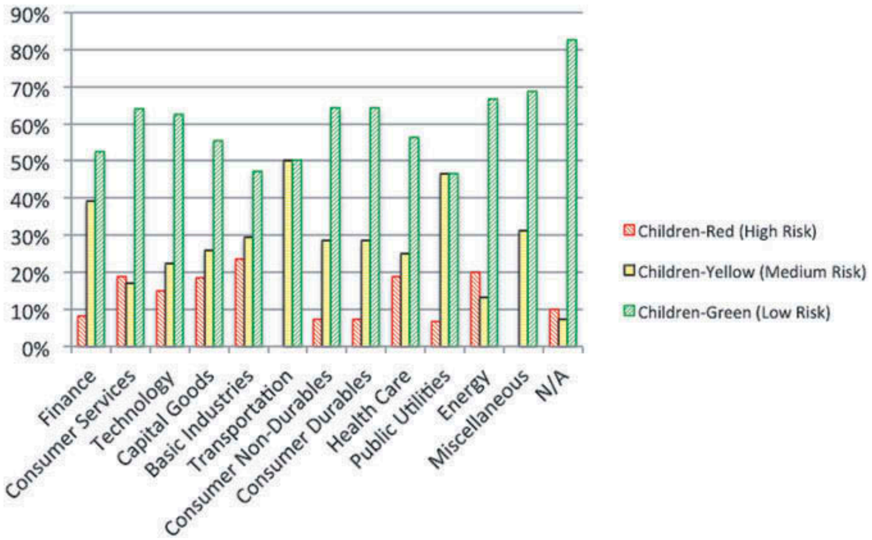

**Figure 7.** Risk level: Collecting PII of children.

### *Data aggregation*

Finally, as Figure 11 demonstrates, almost every company aggregates data. The majority (65%), however, anonymizes the data first. Manual investigation of policies showed that the purpose of data aggregation is usually internal (e.g., to improve their website).

## Conclusions

We studied privacy policies of 10% of all the North American companies listed on NYSE, Nasdaq, and AMEX stock markets. We manually assigned green/yellow/red risk levels for how the policy treats any of the following 10 privacy pertinent factors: E-mail, Credit Card Number, Social Security Number, Ads

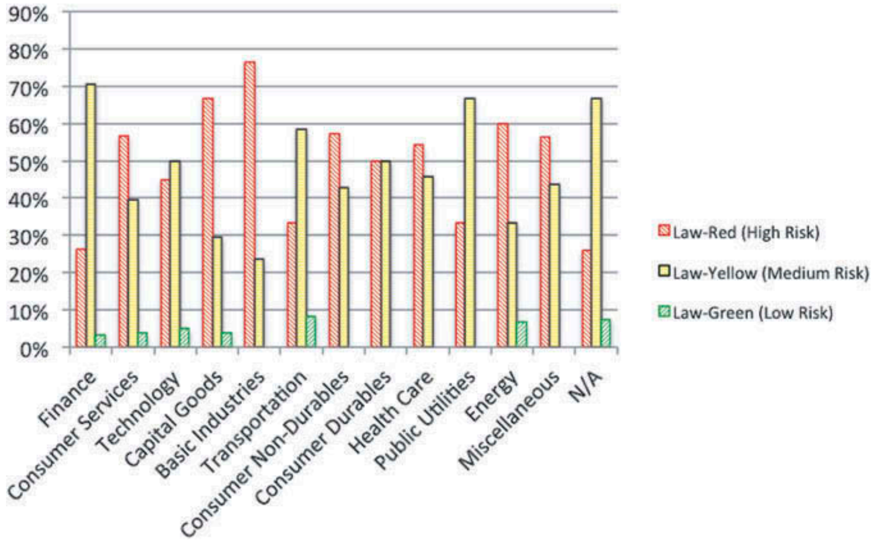

**Figure 8.** Risk level: Sharing with law enforcement.

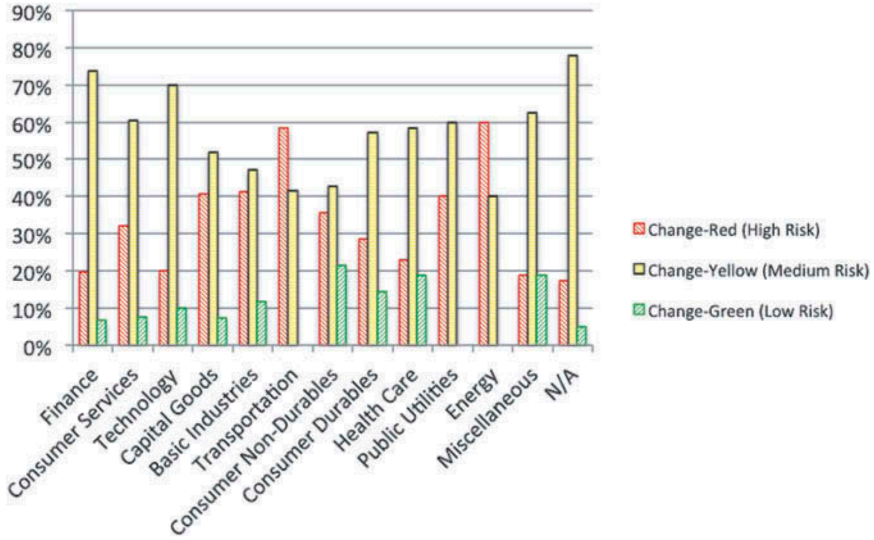

**Figure 9.** Risk level: Policy change notice.

and Marketing, Location, Children, Sharing with Law Enforcement, Notice, Choice, and Aggregation. The study revealed interesting statistics in each of the ICB industries as well as overall. Most importantly, we saw an inclination to collect users' PII but to use only for the expected service of the company. These statistics can assist companies in advancing their privacy practices, regulators in judging the effectiveness of related laws, and users in raising their awareness. We found that:

(1) Strikingly, 31% of these companies do not have any form of privacy policy or notice on their websites.
(2) The companies that did post a privacy policy showed a consistent inclination to toe the line—playing it safe so as to minimize their risk, while simultaneously choosing to gather personal

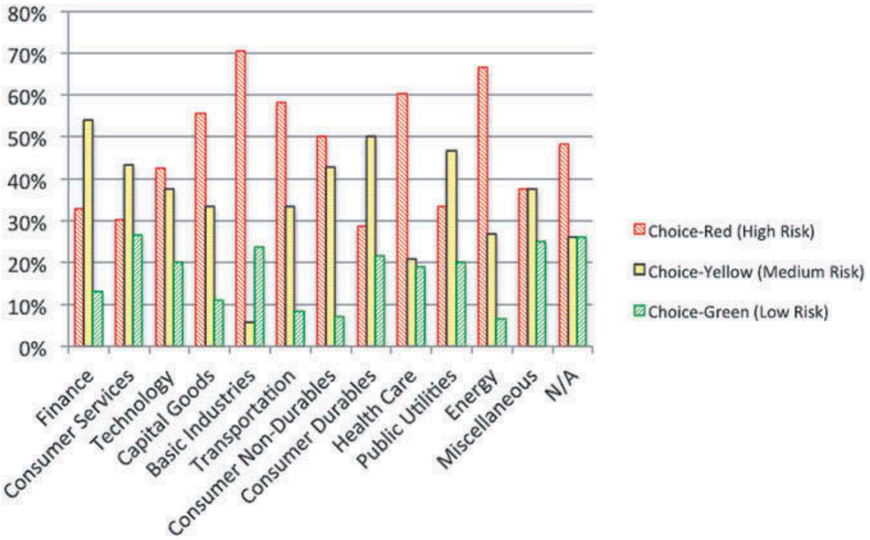

**Figure 10.** Risk level: Choice (Control) of data.

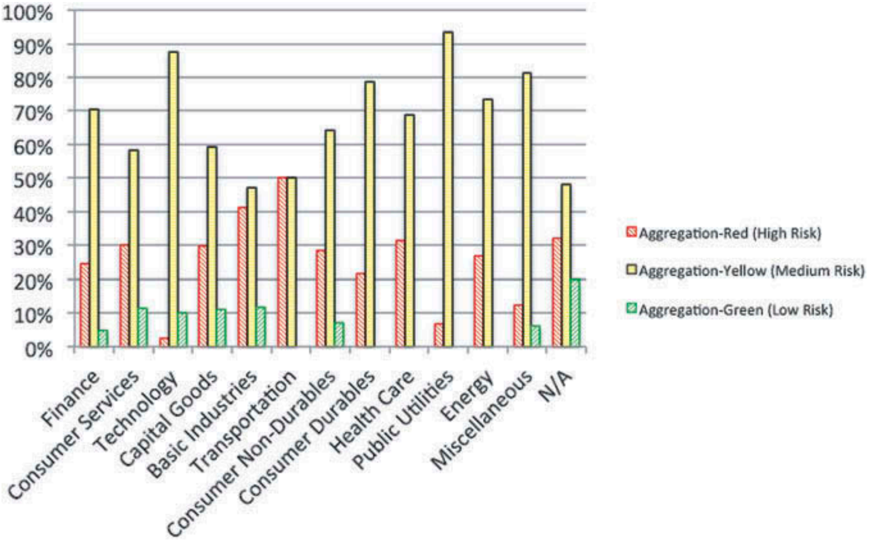

**Figure 11.** Risk level: Aggregation.

information to increase their utility and value. For instance, 81% of the companies (ranging from 53% to 93% across industries) collect e-mail addresses and 64% of the companies use PII (including e-mail addresses) to promote their own services or products.

(3) Regulation, e.g., protecting children's PII (COPPA, 1998), has positively affected privacy policies with respect to children (FTC, 2002): as little as 13% of policies studied collect PII of children under 13.

(4) The companies provide users' PII to law enforcement, and 45% do not even ask for official documents like a warrant or subpoena.

(5) The majority (63%) expect the user to frequently check the posted privacy policy and consider the continued use of the website as implicit agreement to any changes in the privacy policy.
(6) Many companies allow users to edit or correct PII gathered about them, but, surprisingly, one out of three (35%) does not let users completely delete their records.
(7) The majority (65%) of the companies studied aggregate and anonymize data by taking out PII and use the aggregated data for internal purposes such as improving their website.
(8) Even though some industries are particularly lacking with respect to privacy practices (e.g., 50% of the companies in the general Energy industry lack privacy policies), the trend remains the same across all industries.

### *Example application for researchers and users: privacycheck*

In addition to the above conclusions, our study produced a comprehensive annotated corpus of privacy policies. In order to show the applicability of the corpus produced through this study, we used this corpus as the training set for the data mining models that enable PrivacyCheck. PrivacyCheck is a browser extension developed at the Center for Identity at the University of Texas at Austin, currently freely available for Google Chrome and Mozilla Firefox (Nokhbeh Zaeem et al., 2015; UT CID, 2015) that gives users a quick and easy to understand overview of 10 important factors discussed in privacy policies. When the user provides the URL of the company's privacy policy page, PrivacyCheck accesses the text of the page using a data mining algorithm. PrivacyCheck automatically summarizes that page, returning icons that indicate the level of risk (green/yellow/red) for the 10 factors we investigated in this paper. PrivacyCheck is currently installed on 436 Chrome browsers. PrivacyCheck shows how the corpus of privacy policies is useful for future research as an annotated corpus as well as for end-users as a tool.

### *Example application for companies*

Many small businesses pick a privacy policy from a default list of options available on the Internet. While it is useful for small businesses to employ a default privacy policy, they might be unaware of how the default privacy policy they picked without understanding the technical details compares to other businesses in their industry. Our work sheds light on the status quo of privacy policies in each industry, which can benefit small and large businesses alike. Furthermore, our results are of particular benefit to small businesses, which may lack the capacity or knowledge to independently assess privacy policies in their industry.

### *Example application for regulators*

Continued assessment of laws and regulations reveals how regulators have been successful and paves the way for future regulation. Our work is useful in evaluating the effect of previous regulations as well as setting a baseline for future regulations. For example, we found that, the Children's Online Privacy Protection Act (COPPA) (COPPA, 1998) has positively affected privacy policies (FTC, 2002) and limited the number of websites that collect PII of children under 13 to about 13%.

## Acknowledgments

This work was supported in part by the State of Texas IDWise Project. The authors would like to thank the Texas Legislature for its vision and investments in research advances and technology to support consumer empowerment in understanding and managing their online privacy and identity assets.

This work was also funded in part by the Center for Identity's Strategic Partners. The complete list of Partners can be found at https://identity.utexas.edu/strategic-partners.

The authors would also like to thank Rachel L. German, Raphael De Los Santos, Usman Mahmood, Ali Ziyaan Momin, Blake Muir, and Haoran Niu for reading privacy policies, and James E. Zaiss for proofreading this paper.

## Notes on contributors

***Razieh Nokhbeh Zaeem*** received her Ph.D. in Electrical and Computer Engineering from the University of Texas at Austin in 2014. In 2010, she was honored as a Google Anita Borg Scholarship Finalist. She interned at Rockwell Automation Inc. in Austin, TX in 2010, and at Fujitsu Laboratories of America in Sunnyvale, CA in 2012. She joined the Center for Identity as a post-doctoral fellow in 2014 and is now a research scientist at the Center. She has published in prestigious journals and conferences on a broad range of topics from automated software engineering and data mining to privacy concerns and identity protection.

***Suzanne Barber*** is the AT&T Endowed Professor in Engineering in the Department of Electrical and Computer Engineering and Director of the Center for Identity at The University of Texas at Austin. Previously serving as the Director of Software Engineering at The University of Texas at Austin, Dr. Barber led the cross-disciplinary Center for Excellence in Distributed Global Environments (EDGE).

## ORCID

Razieh Nokhbeh Zaeem http://orcid.org/0000-0002-0415-5814

## References

Bhatia, J., Breaux, T. D., & Schaub, F. (2016). Mining privacy goals from privacy policies using hybridized task recomposition. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *25*(3), 22. doi:10.1145/2907942

Brown, D. H., & Layne Blevins, J. (2002). The safe-harbor agreement between the United States and Europe: A missed opportunity to balance the interests of e-commerce and privacy online? *Journal of Broadcasting & Electronic Media*, *46*(4), 549–564. doi:10.1207/s15506878jobem4604_5

Cha, J. (2011). Information privacy: A comprehensive analysis of information request and privacy policies of most-visited web sites. *Asian Journal of Communication*, *21*(6), 613–631. doi:10.1080/01292986.2011.615942

COPPA. (1998). *Coppa: Children's online privacy protection act*. Retrieved from http://www.coppa.org

Cranor, L. F., Egelman, S., Sheng, S., McDonald, A. M., & Chowdhury, A. (2008). P3p deployment on websites. *Electronic Commerce Research and Applications*, *7*(3), 274–293. doi:10.1016/j.elerap.2008.04.003

Cranor, L. F., Leon, P. G., & Ur, B. (2016, August). A large-scale evaluation of U.S. financial institutions' standardized privacy notices. *ACM Transactions Web*, *10*(3), 17:1–17: 33. doi:10.1145/2911988

Culnan, M. J. (1999). *Georgetown Internet privacy policy survey: Report to the federal trade commission*. Washington, DC: Georgetown University, The McDonough School of Business.

Disconnect Me. (2014). *disconnect me privacy icons*. Retrieved from https://disconnect.me/icons

FTC. (1998). *Privacy online: A report to congress*. Retrieved from https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-report-congress/priv-23a.pdf

FTC. (2000). *Privacy online: Fair information practices in the electronic marketplace: A federal trade commission report to congress*. Retrieved from https://www.ftc.gov/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission

FTC. (2002). *Protecting children's privacy under COPPA: A survey on compliance*. Retrieved from https://www.ftc.gov/reports/protecting-childrens-privacy-under-coppa-survey-compliance

FTC. (2010). *Exploring privacy: An FTC roundtable discussion*. Retrieved from https://www.ftc.gov/sites/default/files/documents/public_events/exploring-privacy-roundtable-series/privacyroundtable_march2010_transcript.pdf

FTC. (2012). *Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers*. Retrieved from https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers

Graber, M. A. D., Alessandro, D. M., & Johnson-West, J. (2002). Reading level of privacy policies on internet health web sites. *Journal of Family Practice*, *51*(7), 642–642.

ICB. (2006). *Industry classification benchmark (ICB): A single standard defining the market*. Retrieved from http://www.icbenchmark.com

Kohavi, R. (2001). Mining e-commerce data: The good, the bad, and the ugly. *In International Conference on Knowledge Discovery and Data Mining*, 8–13.

Li, Y., Stweart, W., Zhu, J., & Ni, A. (2012). Online privacy policy of the thirty Dow Jones corporations: Compliance with FTC fair information practice principles and readability assessment. *Communications of the IIMA*, *12*(3), 5.

Liu, C., & Arnett, K. P. (2002). Raising a red flag on global www privacy policies. *Journal of Computer Information Systems*, *43*(1), 117–127.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, *15*(4), 336–355. doi:10.1287/isre.1040.0032

McDonald, A. M., & Cranor, L. F. (2008). Cost of reading privacy policies, the. *I/S: A Journal of Law and Policy for the Information Society*, *4*, 543.

Meinert, D. B., Peterson, D. K., Criswell, J. R., & Crossland, M. D. (2006). Privacy policy statements and consumer willingness to provide personal information. *Journal of Electronic Commerce in Organizations*, *4*(1), 1. doi:10.4018/jeco.2006010101

Milne, G. R., & Culnan, M. J. (2004). Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. *Journal of Interactive Marketing*, *18*(3), 15–29. doi:10.1002/dir.20009

Milne, G. R., Culnan, M. J., & Greene, H. (2006). A longitudinal assessment of online privacy notice readability. *Journal of Public Policy & Marketing*, *25*(2), 238–249. doi:10.1509/jppm.25.2.238

Nasdaq. (2015). *Nasdaq*. Retrieved from http://www.nasdaq.com

Nokhbeh Zaeem, R., German, R. L., & Barber, K. S. (2015). Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology*. (Submitted)

Pan, Y., & Zinkhan, G. M. (2006). Exploring the impact of online privacy disclosures on consumer trust. *Journal of Retailing*, *82*(4), 331–338. doi:10.1016/j.jretai.2006.08.006

Rains, S. A., & Bosch, L. A. (2009). Privacy and health in the information age: A content analysis of health web site privacy policy statements. *Health Communication*, *24*(5), 435–446. doi:10.1080/10410230903023485

Regard, H. (1980). *Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data*, OECD.

Romanosky, S., Telang, R., & Acquisti, A. (2011). Do data breach disclosure laws reduce identity theft? *Journal of Policy Analysis and Management*, *30*(2), 256–286. doi:10.1002/pam.20567

Ryker, R., Lafleur, E., McManis, B., & Cox, K. C. (2002). Online privacy policies: An assessment of the fortune e-50. *Journal of Computer Information Systems*, *42*(4), 15–20.

Shalhoub, Z. K. (2006). Content analysis of web privacy policies in the GCC countries. *Information Systems Security*, *15*(3), 36–45. doi:10.1201/1086.1065898X/46183.15.3.20060701/94186.6

Storey, V. C., Kane, G. C., & Schwaig, K. S. (2009). The quality of online privacy policies: A resource- dependency perspective. *Journal of Database Management*, *20*(2), 19. doi:10.4018/jdm.2009040102

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. doi:10.1287/isre.1090.0260

Turow, J., & Center, A. P. P. (2003). *Americans & online privacy: The system is broken*. Philadelphia, PA, USA: Annenberg Public Policy Center, University of Pennsylvania.

Usable Privacy. (2016). *Usable privacy project website*. Retrieved from https://usableprivacy.org/

UT CID. (2015). *Privacycheck*. Retrieved from https://chrome.google.com/webstore/detail/privacycheck/poobeppenopkcbjejfjenbiepifcbclg

Wilson, S., Schaub, F., Dara, A., Cherivirala, S. K., Zimmeck, S., Andersen, M. S., … Sadeh, N. (2016). Demystifying privacy policies using language technologies: Progress and challenges. In *TA-COS '16: LREC workshop on text analytics for cybersecurity and online safety*.

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., et al (2016). The creation and analysis of a website privacy policy corpus. *Annual meeting of the association for computational linguistics* (pp. 1330–13340).

Zahir Irani, P. C., Sipior, J. T., Ward, B., & Connolly, R. (2013). Empirically assessing the continued applicability of the IUIPC construct. *Journal of Enterprise Information Management*, *26*(6), 661–678. doi:10.1108/JEIM-07-2013-0043

Zhang, X., Toru, S., & Kennedy, M. (2007). A cross-cultural analysis of privacy notices of the global 2000. *Journal of Information Privacy and Security*, *3*(2), 18–36. doi:10.1080/15536548.2007.10855814

## Appendix

Online survey of important privacy pertinent factors

We want to know what parts of a privacy policy users care most about. This form is designed to collect your feedback on what you care about. Answer the following questions on a scale of 1 to 4, with 1 being “do not care” and 4 being “care a great deal”. Please try to discriminate between the items that are most important and those that are somewhat important. Limit responses of “care a great deal” to those items that you feel are most important to keep private.

***The information that you enter when interacting with a website***

How much do you care about the way that a website deals with your…

- Name
- E-mail address
- Phone number
- Billing information (credit card number)
- Social Security Number
- Driver’s License Number
- Personal health information, employer or health care plan information
- Education and work history
- Personally Identifiable Information of your (under 13-year-old) child

***The information that a website collects automatically***

How much do you care if a website gathers and uses information about your…

- Device and software data, for example device type, operating system, browser type and version, browser plug-in types and versions, IP address, MAC address, time zone setting, and screen resolution
- Cookies, for example cookie number, and Flash cookies (also known as Flash Local Shared Objects)
- Viewed or searched products
- Purchase history and credit history information from credit bureaus
- Browsing pattern, for example URL click stream to/through/from their website, page response times, download errors, length of visits to pages, page interaction information (such as scrolling, clicks, and mouse-overs)
- Social networking accounts
- Login and password for other websites

***The information that a website can collect when you are on a mobile device***

How much do you care about the way that a website deals with your…

- Exact location

***Usage***

How much do you care if the website uses any of the information mentioned above for…

- Processing orders for products or services, and responding to questions
- Improving customer services
- Delivering personalized content within the site, providing search results and links (including paid listings and links)
- Ads, marketing, communication regarding updates, offers, and promotions
- Monitoring and ensuring site integrity and security, protecting the rights or safety of other users
- Aggregating non-identifiable information for business analysis
- Complying with the law and governmental requests
- Credit risk reduction, and collecting debt
- Transferring of assets if the company is acquired
- Determining your geographic location, providing location-based services
- Measuring the effectiveness of ads and user interactions with them

***Other***

How much do you care about the website’s policy for…

- Updating their privacy policy
- Allowing you to update or delete your information
- Enforcing the privacy policy
- Retaining data