PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining

Razieh Nokhbeh Zaeem
Rachel L. German
K. Suzanne Barber

2016

UT CID Report #1602
Research shows that only a tiny percentage of users actually read the online privacy policies we all implicitly agree to when using a website. It also suggests that users ignore privacy policies because they are lengthy and, on average, require two years of college education to comprehend. We propose a novel technique that tackles this problem by automatically extracting graphical summaries of online privacy policies. We use data mining models to analyze the text of privacy policies and answer ten basic questions concerning the privacy and security of user data, what information is gathered from them, and how this information is used.

In order to train the data mining models, we thoroughly study privacy policies of 400 companies (7% of all listings on NYSE, Nasdaq, and AMEX stock markets) across industries. Our free Chrome browser extension, PrivacyCheck, utilizes the data mining models to summarize any HTML page that contains a privacy policy. PrivacyCheck stands out from currently available counterparts because it is readily applicable on any online privacy policy. Experimental results show that PrivacyCheck summaries are accurate 60% of the time. Over 350 independent Chrome users are currently using PrivacyCheck.

Categories and Subject Descriptors: H.2.8 [Database Applications]: Data mining; K.4.1 [Public Policy Issues]: Privacy

General Terms: Legal Aspects

Additional Key Words and Phrases: Privacy policy, data mining, classification

ACM Reference Format:
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

When consumers give Personally Identifying Information (PII) on the Internet, they often have no idea what companies will do with it. Federal and state laws require most businesses to publicly post a privacy policy stating how they use users’ PII. The Federal Trade Commission (FTC) has successfully prosecuted companies for deceptive and misleading practices when using personal data in ways contrary to their stated privacy policies. As a result, most companies have privacy policies that are easily accessible online. The problem is that many people never read these lengthy and often technical documents.

This work is supported by the State of Texas IDWise Project. This work was in part funded by the Center for Identity’s Strategic Partners. The complete list of Partners can be found at https://identity.utexas.edu/strategic-partners.

Author’s addresses: R. Nokhbeh Zaeem, R. L. German, and K. S. Barber, Center for Identity, University of Texas at Austin, {razieh, rachel, sbarber}@identity.utexas.edu.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2010 ACM 1539-9087/2010/03-ART39 #15.00
DOI: http://dx.doi.org/10.1145/0000000.0000000

ACM Transactions on Internet Technology, Vol. 9, No. 4, Article 39, Publication date: March 2010.
Research suggests that while most users know about privacy policies, less than half of them have ever read a privacy policy [Meinert et al. 2006]. Studies that used self-reported data from users found that only 4.5% claim to always read them [Milne and Culnan 2004]. However, the more reliable server side observation of websites reveals even more astonishing statistics that only 1% or less of users click on a website’s privacy policy [Kohavi 2001]. More recent studies using advanced eye tracking techniques show that the same still holds true today: users barely take effort to read privacy policies thoroughly [Steinfeld 2016].

This is, in part, attributed to the fact that it is often difficult to read and comprehend privacy policies. A study of the readability of privacy policies showed that the average privacy policy required two years of college level education to comprehend [Graber et al. 2002; Milne et al. 2006]. An analysis of 80 privacy policies for top health websites found that none of the websites had a privacy policy that was comprehensible by most English-speaking individuals in the United States [Graber et al. 2002]. In addition, a study of online privacy policies found that privacy policies are getting longer and harder to read, with the readability of the privacy policies decreasing over time, and the average length of the policies increasing by more than 500 words [Milne et al. 2006]. In fact, reading privacy policies is so time consuming that if users were to read each new privacy policy they encounter in a year, it would take them over 200 hours [McDonald and Cranor 2008].

PrivacyCheck is a valuable Privacy Enhancing Technology (PET) that gives users a quick and easily understood overview of the essential content of a company’s online privacy policy. This browser add-on is intended to provide a graphical, at-a-glance description of the ways in which companies use their users’ personal data. Similar, currently available resources fall short of this goal due to their lack of comprehensive web coverage. TRUSTe [TRUSTe 2015], P3P [W3C 2006], and Nutrition Label [Kelley et al. 2010; Kelley et al. 2009], for example, require businesses to enroll in a sign-up service or use a new format. Likewise, many other crowd sourced efforts, for example ToS;DR [ToS;DR 2012], only provide information for websites that have been manually checked. PrivacyCheck does not have these limitations and surpasses existing add-ons, apps, and certifications by accessing the text of any web page using a data mining algorithm. When the user provides the URL of the company’s privacy policy page, PrivacyCheck searches that page, returning icons that indicate the level of risk for several factors that impact the security and privacy of a user’s identity.

PrivacyCheck summarizes a privacy policy with respect to a list of ten privacy factors (Section 2.1). For each of these factors, e.g., email address, PrivacyCheck answers a basic question, e.g., “How does the site handle your email address?” The answers to these questions are mapped to three levels of risk: red (high risk), yellow (medium risk), and green (low risk). For example, if a website does ask for users’ email addresses, but states in the privacy policy that uses them only for the intended service, it is ranked at the yellow risk level with respect to this PII factor. We surveyed privacy experts and utilized previous research to compile the list of factors.

PrivacyCheck automatically predicts risk values for each privacy factor using a classification data mining (supervised machine learning) model. Our key insight in developing PrivacyCheck was to train a data mining model against privacy policies for each factor, and then use it to predict risk values for the factor when checking new privacy policies. To train the models, a team of seven privacy experts, graduate and undergraduate students read 400 privacy policies randomly selected from the NYSE, Nasdaq and AMEX company listings, and manually assigned risk levels to the ten factors.

In this paper, we make the following contributions:

— We use data mining to summarize privacy policies online.

ACM Transactions on Internet Technology, Vol. 9, No. 4, Article 39, Publication date: March 2010.
We present PrivacyCheck, a browser extension that readily extracts risk levels for privacy factors from privacy policies and presents them in a graphical way.

We carefully investigate 400 privacy policies (from 7% of all companies listed on NYSE, Nasdaq, and AMEX stock markets) and use them to train data mining models.

We evaluate PrivacyCheck on 50 other policies and show how it exceeds similar tools and certifications.

This paper is organized as follows. Section 2 explains the user interface and technical foundation of PrivacyCheck. Section 3 puts PrivacyCheck in context by surveying other available PET tools and services. Section 4 evaluates PrivacyCheck and compares it to the other tools discussed in Section 3. Finally, Section 5 outlines some of the use cases of PrivacyCheck and Section 6 concludes the paper.

2. PRIVACYCHECK BROWSER EXTENSION

PrivacyCheck is currently implemented as a browser extension for Google Chrome and is publicly available [CID 2015]. Figure 1 shows a snapshot of the PrivacyCheck browser extension. The user first navigates to the URL of a privacy policy and then opens the browser extension and clicks its start button. PrivacyCheck extracts the text of the privacy policy, pre-processes it, and sends it to the data mining model to determine the level of risk for each of the ten privacy factors. PrivacyCheck then displays the risk levels as red (high risk), yellow (medium risk), and green (low risk), which are more elaborately explained once the user hovers over each item (as seen in Figure 1).

2.1. PrivacyCheck Questions

The ten questions that PrivacyCheck answers are:

1. How does the site handle your email address?
2. How does the site handle your credit card number and home address?
Table I. Risk Levels for Privacy Factors.

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

Table I shows the risk levels for each of the privacy factors. The red risk level would be assigned in case the information about any of the factors is not disclosed.

In order to choose the questions that PrivacyCheck seeks to answer, we evaluated related work and performed a survey.


FTC recommends that privacy policies follow Fair Information Practice Principles (FIPP’s) [FTC 2000]: Notice, Choice, Access, Security, and Enforcement. We also reviewed public submissions and staff reports from several workshops and round-tables that FTC held in 2010 and 2012 [FTC 2010; 2012] that suggested these privacy factors: Aggregation, Encryption, Third Party Sharing, Sharing with Law Enforcement, Security, Access, Control, Usage, Ads, Retention, and Location.


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.

2.1.2. 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 and graduate student employees 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 appendix A. Using the results of the survey, we enlisted the factors that PrivacyCheck evaluates and the questions it answers about them. We manually assigned the answers and risk levels by carefully reviewing the privacy policies used for training (Section 2.3).

2.2. Architecture

Figure 2 shows a high level architecture of the PrivacyCheck extension. The browser client (written in HTML and Java-script) initially determines whether the given URL indeed points to a privacy policy. In order to do so, it follows the algorithm explained in Section 2.4 to get the related text. It then sends the related text to the application server (written in Python) which relays the text to the data mining server. The data mining server checks the text against a trained classification model to find out if this is a privacy policy. The result of the classification model is sent back to the application server which then forwards it to the browser extension. The browser extension checks the result and, if not a privacy policy, alerts the user that the URL does not point to a privacy policy. If the URL is determined to be for a privacy policy, however, the browser extension follows the same algorithm (Section 2.4) to extract related paragraphs for each privacy factor. The related text snippets for privacy factors are asynchronously sent to the application server and then the data mining server. The data mining server has a classification model trained for each factor (Section 2.5), which it uses to classify the text snippet according to the risk levels (high, medium, and low). Once the browser extension client receives the result for a factor, it shows them as red, yellow, and green for high risk, medium risk, and low risk.

The intermediate application server is put in place to make encrypted connections to the data mining server. The application server uses a RESTful API to receive requests for a specific privacy factor and provided text from the browser extension. It uses its private key with read permissions to connect to the data mining server. The private key should not be included in the available source code of the client, to avoid unwanted requests to the paid data mining server.

2.3. Corpus of 400 Privacy Policies

We compiled a set of 400 privacy policies that we use for several purposes: collecting keywords (Section 2.4), designing answers and risk levels (Section 2.1), and training the classification models (Section 2.5).

A scientific methodology of selecting companies is central to collecting a comprehensive and generalizable corpus of their privacy policies. We aimed for a selection of companies that:

1. Were reputable companies, i.e. listed by well known industrial entities.
2. Were categorized based on a standard and commonly used industrial classification.
3. 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 (Section 2.3.1) using Industry Classification Benchmark (Section 2.3.2).

2.3.1. NYSE, Nasdaq and AMEX. The New York Stock Exchange (NYSE), Nasdaq, and the American Stock Exchange (AMEX) are American stock exchange markets, and are respectively the first, second, and third largest stock exchange by market capitalization in the US. The Nasdaq Company List includes companies listed on Nasdaq, as well as NYSE and AMEX. As of the date of this paper, the companies listed by these three stock markets add up to 6,500 companies worldwide, most of them in North America [Nasdaq 2015].

We used this company list for training of PrivacyCheck. The company list uses Industry Classification Benchmark and includes the name, the ICB industry of the company, and the link to its websites.

2.3.2. Industry Classification Benchmark. Industry Classification Benchmark (ICB) is an industry classification taxonomy. It segregates markets into 10 industries which are in turn partitioned into 114 sub-sectors. Each company is allocated to a sub-sector that most closely resembles its majority source of revenue [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].

2.3.3. Company Selection. 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 paper. We randomly selected 400 companies, evenly across industries, covering 7%. Once we had the list of companies to study, we first found the URL 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 to the company’s privacy policy. All the 400 links were manually checked to make sure they point to the latest version of the company’s privacy policy.

2.4. PrivacyCheck Client Side: Text Pre-processing

We utilized a text pre-processing algorithm to extract parts of privacy policies that are related to each of the privacy factors. Algorithm 1 breaks the text into paragraphs, removes punctuation, converts uppercase to lowercase, removes stop words, and finally keeps only the paragraphs that have at least one keyword related to a particular factor.

The input to this algorithm is the text from the web page \(T\), a set of privacy factors \(F\) that we would like to consider, and the set of keywords \(K_f\) for each factor \(f\) (Table II elaborates on these keywords). The algorithm’s output is a text snippet \(S_f\) for each of these factors. Line 1 breaks the web page text into paragraphs. Then, for each paragraph, the algorithm performs the following. Line 3 replaces all non alphanumeric characters with spaces, effectively removing all the punctuation marks to create the punctuation-less paragraph \(pl\). Line 4 converts this punctuation-less paragraph to lowercase \(lc\). The next line removes all the stop words (stop word-less \(sl\)), and finally Line 6 replaces any sequence of spaces (generated through previous text manipulating lines or originally present in the text) with one single sequence. This final paragraph \(fp\) has only those words that carry meaning for the purpose of data mining in lowercase. Line 7 puts those words in \(W\). Line 8 iterates through the factors and for each
factor does the following steps. If any word of this paragraph contains any keyword for the factor then the entire paragraph \(fp\) is kept in \(S_f\) for that factor and the algorithm moves on to the next factor. Finally, after iterating over each factor and over each paragraph, the algorithm returns \(S\) which contains all \(S_f\)'s.

**ALGORITHM 1: Text Pre-processing**

```plaintext
input : Web Page Text \(T\), Set of Privacy Factors \(F\), Set of Keywords for Each Factor \(K_f\)
output: Text Snippet for Each Privacy Factor \(S_f\)

1. \(P \leftarrow T.split(\"\n\")
2. foreach paragraph \(p\) in \(P\) do
   3. \(pl \leftarrow p.replace(/[\^A-Za-z0-9-]/g,"\") // punctuation-less removes any non alphanumeric character
   4. \(lc \leftarrow pl.toLowerCase() // converts to lowercase
   5. \(sl \leftarrow lc.replace(\/\b(i|me|my|\ldots|should|now)\b/g,"\") // stopword-less removes stop words
   6. \(fp \leftarrow sl.replace(\s{2,}/g,"\") // final paragraph replaces any double or more spaces with a single space
   7. \(W \leftarrow fp.split(\"\")
   8. foreach factor \(f\) in \(F\) do
      9. nextFactor:
      10. foreach word \(w\) in \(W\) do
          11. if \(w.contains(k)\) then
              12. \(S_f \leftarrow S_f + fp + '\', break nextFactor
          13. end
      14. end
      15. end
      16. end
      17. end
      18. end
      19. return \(S\)
```

We investigated the 400 privacy policies in order to determine the list of keywords related to each of the privacy factors. The methodology of selecting keywords was largely manual, identifying frequent words in the paragraphs that seemed related to any given privacy factor. Investigating other methodologies and assessing the selection of these keywords is a future avenue of work. The current sets of keywords are displayed in Table II. The table shows only the word stem (i.e., root form) of the keywords. In the actual implementation, we expanded this list of stems to include other forms of the words too. For example, we generated other forms of the word “locate” (e.g., “location”, “locations”, and “locating”) for the location factor. The algorithm keeps any paragraphs that have at least one form of one of the related keywords for a given factor.

In Table II, the first row indicates keywords used to detect a privacy policy, and the rest are for different factors. These keywords are selected considering not only the privacy factor, for example email address, but also the questions we seek to answer about each factor to assign risk levels, e.g., whether email addresses are shared with third parties. Hence keywords like third party, share, and sell are included for privacy factors that answer similar questions.

2.5. **PrivacyCheck Server Side: Data Mining Models**

Once text-preprocessing is complete, PrivacyCheck sends each text snippet (through the application server) to the data mining server. We trained 11 data mining models,
one for detecting if the corresponding page is a privacy policy and one each for the ten factors.

To train the models, we leveraged the 400 privacy policies. A team of seven privacy experts, graduate and undergraduate students read each of these policies, totaling to close to 700K words, and scored each policy according to Table I using the red/yellow/green levels. We performed quality control by assigning every policy to two team members and comparing and resolving discrepancies for the first %15 of privacy policies (60 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 paper.

Then, we used Algorithm 1 to pre-process the 400 policies and put together a training file for each factor that includes the corresponding text snippet (only those paragraphs that have a related keyword) and the risk level for each of the 400 policies. In order to train the model to determine whether or not the web page is indeed a privacy policy, we utilized the 400 privacy policies with 400 randomly selected web pages that were not privacy policies. We randomly generated these non-privacy-policy pages using a random web page generator\(^1\) and made sure that they indeed are not privacy policies. All the privacy policies underwent the same text-preprocessing using the keywords for Being a Privacy Policy in Table II to compile the training file that includes text snippets and the class (is/is not a policy) for each of the 800 policies.

We uploaded these training files to Google Prediction API [Google 2014a]. We experimented with regression and classification models and found classification models naturally suiting our application and giving better results. Therefore, we trained a classification model against each, the independent variable being the text snippet and the dependent variable being the class (i.e., the risk level). There were two main reasons we chose classification over regression: (1) the accuracy of classification models converged when increasing the number of training policies (going from 50 to 400 in increments of 50) while it did not converge for regression models, and (2) the classification models directly gave classes while a mapping (including a set of thresholds) was needed to get from regression scores to classes.

Figure 3 shows the Classification Accuracy of the models when trained against different numbers of privacy policies selected from the corpus. The model Classification Accuracy (CA) is a number between 0 and 1 reported when training a model on Google Prediction, where 1 is 100% accurate. This is an estimate, based on the amount and quality of the training data, of the estimated prediction accuracy of each model. As this figure shows, the CA of most of the models converges as the number of privacy poli-

\(^1\)http://www.randomwebsite.com

Table II. Keywords for Privacy Factors.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0) Being a Privacy Policy</td>
<td>privacy, policy</td>
</tr>
<tr>
<td>(1) Email Address</td>
<td>email, mail, third, party, share, sell, promote, affiliate</td>
</tr>
<tr>
<td>(2) Credit Card Number</td>
<td>credit, card, bill, debit, pay, third, party, share, sell, promote, affiliate</td>
</tr>
<tr>
<td>(3) Social Security Number</td>
<td>social, security, number, ssn, third, party, share, sell, promote, affiliate</td>
</tr>
<tr>
<td>(4) Ads and Marketing</td>
<td>ad, market, third, party, share, sell, promote, affiliate</td>
</tr>
<tr>
<td>(5) Location</td>
<td>locate, geo, mobile, gps, third, party, share, sell, promote, affiliate</td>
</tr>
<tr>
<td>(6) Collecting PII of Children</td>
<td>age, child</td>
</tr>
<tr>
<td>(7) Sharing w/ Law Enforcement</td>
<td>law, regulate, legal, government, warrant, subpoena, court, judge</td>
</tr>
<tr>
<td>(8) Policy Change</td>
<td>notice, change, update, post</td>
</tr>
<tr>
<td>(9) Control of Data</td>
<td>choice, edit, delete, limit, setting, account, access, update</td>
</tr>
<tr>
<td>(10) Data Aggregation</td>
<td>aggregate, non-identifiable</td>
</tr>
</tbody>
</table>
cies used for training reaches 350, justifying our decision of training models against a corpus of 400 policies.

3. RELATED WORK

In this section, we review tools, services, and privacy enhancing technologies that help users protect their privacy, without having to read privacy policies in detail, in the following categories:

(1) Privacy seals require web page operators to enroll in order to evaluate their privacy policies.
(2) New formats encourage web page operators to adopt machine-readable notation to be automatically interpreted.
(3) Crowd sourced services have an online community that reads and rates privacy policies.
(4) Data mining tools leverage machine learning and natural language processing to semi-automatically annotate privacy policies.
(5) Tracking monitors observe web pages in action, instead of investigating their privacy policies.

3.1. Privacy Seals

Privacy seals are logos of organizations or agencies that evaluate and rate privacy policies. For example, TRUSTe [TRUSTe 2015] is a data privacy management company that examines privacy policies and helps businesses align their privacy policies with legal requirements. While the information provided by TRUSTe is manually extracted, TRUSTe and other similar services suffer from two major drawbacks: (1) they significantly lack comprehensive web coverage; even though TRUSTe owns 67.94% of the market share among all similar services it covers only roughly 55,000 in one million

ACM Transactions on Internet Technology, Vol. 9, No. 4, Article 39, Publication date: March 2010.
Another privacy seal is provided by Better Business Bureau (BBB) [BBBOnline 2015], a non-profit organization that provides free business reviews. However, BBB accredited businesses pay a fee for accreditation review.

Overall, researchers have expressed concerns with privacy seals in general: insufficient scrutiny of privacy seal organizations, negative self-selection of websites that participate in a seal, and users’ illiteracy about privacy seals [Kobsa 2007].

3.2. New Formats
The Platform for Privacy Preferences Project (P3P) [W3C 2006] is a standard for websites to express their privacy policies in a both human and machine readable format. Following such standards enables automatic interpretation of privacy policies. P3P and similar standards require web page operators to adopt new formats. As a result, P3P has always suffered from lack of industry participation. Consequently, the P3P working group was closed in 2006 and P3P 1.1 was never finalized [Cranor 2012].

Currently, P3P is used in only 70,000 in one million web pages [BuiltWith 2015]. The failure of P3P in attracting industry participation, however, is not limited to the number of websites that do not support it. A large fraction of the websites that support P3P chose to include a misrepresenting and minimal version of their privacy policy just to prevent Internet Explorer, the major web browser supporting P3P, from blocking their cookies. In fact, thousand of websites were found to use an identical erroneous policy recommended by a Microsoft support website to avoid cookie blocking by Internet Explorer [Cranor 2012].

Researchers have attempted visualizing privacy policies represented in P3P (e.g., Privacy Bird extension for Internet Explorer [AT&T 2002; Cranor et al. 2006]), some with only little success in improving the comprehension [Reeder et al. 2008]. Internet Explorer 6.0 or later includes a feature that textually presents privacy policies formatted as P3P.

Nutrition Label [Kelley et al. 2010; Kelley et al. 2009; Cranor 2012] introduced another new format that asks website operators to consolidate their privacy policies in a one page standardized format inspired by food and drugs nutrition facts panel. Complying with this new format too places an additional, unwanted burden on website operators.

3.3. Crowd Sourced Services
Terms of Service; Didn’t Read (ToS;DR) [ToS;DR 2012] is a free software project that started in 2012 to address the problem that very few users actually read the terms of service for websites they use. In this project, an online community of volunteers read, discuss, and rate privacy policies. The ratings and discussions are available online and as free software in the form of browser extensions for Mozilla Firefox, Google Chrome, Apple Safari, and Opera. Even though privacy policies addressed in this project are read and rated by humans and discussed thoroughly, the in depth and occasionally long discussion poses a new challenge to the usefulness of the ratings: one might read, although selectively, the original privacy policy instead. Furthermore, the coverage of ToS;DR is even more limited than privacy seals and new formats; only 66 privacy policies are rated so far [ToS;DR 2012].

3.4. Data Mining Tools
Building on ToS;DR, Privee [Zimmeck and Bellovin 2014] combines crowd sourcing with rule and machine learning classifiers to classify privacy policies that are not already rated in the crowd sourcing repository. However, the performance of Privee was
found to be limited by the ambiguity of the natural language. Privee most closely resembles our work, however, (1) it uses a smaller corpus for training (only 66 policies from ToS;DR), (2) its training does not enjoy the consistency of a small team working closely together, and (3) we found it very slow when run on longer privacy policies.

The Usable Privacy Project [Sadeh et al. 2013] takes advantage of natural language processing, machine learning, privacy preference modeling, crowd sourcing, and formal methods to semi-automatically annotate privacy policies. This project [Wilson et al. 2016b] annotates a corpus of policies with attributes and data practices as the first step. The Usable Privacy Project [Ammar et al. 2012] has also used a statistical classifier trained using ToS;DR data to answer two basic questions about privacy policies automatically. While the project plans to release tools that digest information from privacy policies to show in an easy-to-use format, no such tool is available yet.

Similarly, others [Clarke et al. 2012] have proposed semi automated extraction of privacy policy features using crowd sourcing, natural language processing, and privacy preference modeling.

3.5. Tracking Monitors
Ghostery [Ghostery 2015] is a software available as free browser extensions for Mozilla Firefox, Google Chrome, Internet Explorer, Opera, and Apple Safari. Ghostery tracks cookies, tags, web bugs, pixels and beacons and notifies the user of their presence as well as the companies that operate them, giving the user the choice to make informed decisions about blocking them. Similarly, Adblock Plus [AdblockPlus 2015] is a free extension that blocks adds and disables tracking. Adblock Plus is available for Android, Google Chrome, Mozilla Firefox, Internet Explorer, Opera, and Apple Safari among other browsers. Ghostery, Adblock Plus, and other similar services are fundamentally different from PrivacyCheck in that they focus on the actions a website takes, instead of the legal privacy policy it posts. Such actions are hopefully aligned with the privacy policy, even though the alignment is not guaranteed. In addition, tracking monitors do not indicate the usage of the information gathered from users.

4. EVALUATION
In this section, we use three different methods to evaluate how accurately PrivacyCheck works in practice. First, we manually investigate what it shows for 50 new privacy policies (Section 4.1). Second, we compare it to some of the alternative tools discussed in Section 3 for those 50 new privacy policies (Section 4.2). Third, we consider the feedback of the user base community currently employing PrivacyCheck as a browser extension (Section 4.3).

4.1. Testing PrivacyCheck
In order to evaluate how well PrivacyCheck summarizes privacy policies, we tested it against 50 new privacy policies not seen in the training phase. To choose privacy policies for testing, we performed a Google search with terms “privacy policy” [Google 2014b] on November 13, 2014 and selected the first 50 non-sponsored search results that we had not used in the training phase. The set of privacy policies included well-known websites (e.g., Google, CNN, Google, Wikipedia) and less-known websites (e.g., Ello, OwnPhones, and Automattic). We thoroughly read each of these new privacy policies and manually determined the risk level for each of the factors. Then, we ran PrivacyCheck and recorded the risk levels it indicated.

PrivacyCheck correctly identifies all of the 50 test websites as privacy policies. Table III compares the Ground Truth (according to what the policy states, and found by reading the policy) with what PrivacyCheck finds automatically. On average, 60% of the times, PrivacyCheck gave the exact correct risk level. We also found that when
PrivacyCheck did not find the exact correct risk level, it was one risk level off in 32% of the times, on average.

4.2. PrivacyCheck vs Other Tools

We consider representative tools from the categories discussed in Section 3. We ignore Privacy Seals and Tracking Monitors, because the former is only the indication of participation in a seal and the latter does not provide static information about a website, rather dynamically tracks it in action. Therefore, we evaluate one New Format (P3P), one Crowd Sourced Tool (ToS;DR), and one Data Mining Tool (Privee) as shown in Table IV.

Since the closest tools to PrivacyCheck discussed in Section 3 are ToS;DR and Privee, we deliberately used top ranking companies through a Google search to give these tools an edge, since the crowd sourcing efforts mostly concentrate on well known companies. Even though only 66 privacy policies are rated in the ToS;DR project, we had 14 policies in the testing phase from its list of reviewed policies to be able to compare the output of PrivacyCheck to it.

ToS;DR provided ratings for 14 test websites, as we selected the websites with a focus on ToS;DR. However, only for 7, a classification was available. For the other 7 websites, even though a classification was not available, thumbs up and down (positive and negative) points were given. Examples of thumbs up points are: (1) users can request access and deletion of personal information, (2) terms and privacy policy pages are organized and formatted well, and (3) the service does not track users at all. Examples of thumbs down points include: (1) terms may be changed any time without notice to the user, (2) the service tracks users on other websites, and (3) the service may sell user data as part of a business transfer.

Privee extension for Chrome [Zimmeck 2014] is supposed to show the exact information as ToS;DR when it is available. However, we found 4 discrepancies where it did not. If ToS;DR does not have a record for a website, Privee uses its machine learning
Table IV. PrivacyCheck vs Other Tools (– Means Results Unavailable, Up and Down Stand for Thumbs Up and Down).

<table>
<thead>
<tr>
<th>Company</th>
<th>PrivacyCheck</th>
<th>P3P</th>
<th>ToS;DR</th>
<th>Privee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slack</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Ello</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>CNN</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Lego</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>PNC</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Hilton</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>Monster</td>
<td>2</td>
<td>8</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>National Weather</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>Sony Music</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>EA</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>BlackBerry</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Dep of Justice</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>Ziff Davis</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>AdRoll</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>UT Austin</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Microsoft</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>70 Lines</td>
</tr>
<tr>
<td>RoyalAirForce</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>OwnPhones</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Twitter</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>Wordpress</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Google</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>SoundCloud</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>–</td>
</tr>
<tr>
<td>GitHub</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>TwitPic</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Delicious</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Dell</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Apple</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Walmart</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Staples</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>500Pics</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>MyKolab</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>IBM</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>Facebook</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>–</td>
</tr>
<tr>
<td>Yahoo</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Automattic</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>–</td>
</tr>
<tr>
<td>Adobe</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>AddThis</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Verison</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>NBCUniversal</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>Ebay</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>RocketFuel</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Disney</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Pandora</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>Zynga</td>
<td>2</td>
<td>8</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>AOL</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>–</td>
</tr>
<tr>
<td>Snapchat</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Pinterest</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>–</td>
</tr>
</tbody>
</table>

ACM Transactions on Internet Technology, Vol. 9, No. 4, Article 39, Publication date: March 2010.
classifiers to learn from crowd sourced information provided by ToS;DR and build on it. The Privee extension labels each policy with an overall letter grade, from A to C, by considering collection, profiling, ad tracking, ad disclosure, retention, and encryption.

We used Internet Explorer 11 to investigate P3P formatted privacy policies. Only one website (Microsoft) provided its policy in the P3P format, which was summarized as a 70 line description by Internet Explorer. The summary included what kind of information is collected and why, who has access to the information, how long it is retained, whether users have access to the information, and how disputes are handled.

In conclusion, PrivacyCheck is applicable on a wider range of privacy policies, compared to its counterparts. We were not able to find a matching between Privee and PrivacyCheck results, however, this might be expected as the two tools use different sets of privacy factors and are trained on different numbers of policies. For this testing experiment, we looked mostly at privacy policies that were covered by ToS;DR, and consequently were from high profile companies. Nonetheless, even those companies rarely supported new formats.

4.3. PrivacyCheck User Base

PrivacyCheck is currently installed on 366 Chrome browsers and has a rating of 4.56 (rated by 9 users) on a scale of 1 to 5 (Figure 4) [CID 2015]. According to the reviews, users found it “working as intended” and “self explanatory”. One problem that users had with PrivacyCheck was that it does not work on privacy policies that do not discuss the ten factors. We feel that this shortcoming can be overcome by investigating more privacy factors. Also, PrivacyCheck does not work on privacy policies in the PDF format.

5. APPLICABILITY

By using PrivacyCheck, consumers will be able to gather fast and informed knowledge of how companies are using their personal information. By doing so, PrivacyCheck
enables consumers to gain concrete awareness of an abstract topic in privacy. Research indicates consumer valuations of privacy are highly susceptible to contextual factors—PrivacyCheck is a step in the direction of making privacy decisions easier and more concrete for consumers.

Many businesses, and particularly 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, being unaware of how they have pledged to manage, store, and use consumer data retains harmful consequences. For instance, a small business who is breached may lack the knowledge of how they should inform the consumers, in a manner that is compatible with the default privacy policy they picked without understanding the technical details. Furthermore, small businesses may lack the technical expertise to understand how the cryptic language in their privacy policy relates to what they intend or want to do with consumer data. By using PrivacyCheck on their own privacy policy, small businesses can better understand their promises to the consumers regarding the handling of personal information. PrivacyCheck can also enable small businesses to better understand their enacted privacy policy, and therefore, better communicate this vision of handling personal information to their consumers. In these ways, PrivacyCheck offers a novel and technically simple way to understand and communicate the details of a privacy policy for both businesses and consumers.

6. CONCLUSIONS

In this paper, we presented a novel data mining-based technique, accompanied with its free implementation as a browser extension, to automatically sum up online privacy policies and show them as graphical icons with short descriptions. We identified, through literature review and a survey of privacy experts, ten essential questions users should ask about how businesses use their PII. Our browser extension, PrivacyCheck, automatically answers these ten questions for any given privacy policy using data mining classification models that are trained on 400 policies and operate on a server. PrivacyCheck assigns a risk level (green, yellow, or red) to the privacy policy for each of the ten factors in question. Unlike the other somewhat similar counterparts we discussed in this paper, PrivacyCheck is readily and universally applicable on privacy policies. We evaluated PrivacyCheck and found that its results are accurate 60% of the time. Finally, PrivacyCheck proved to be useful to an independent body of users, being installed hundreds of times on Google Chrome web browser.

ACKNOWLEDGMENTS

The authors would like to thank the Texas Legislature for its visions and investments in research advances and technology to support consumer empowerment in understanding and managing their online privacy and identity assets.

The authors also would like to thank Raphael De Los Santos, Usman Mahmood, Ali Ziyaan Momin, Blake Muir, and Haoran Niu for reading privacy policies to train PrivacyCheck.

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

REFERENCES


ACM Transactions on Internet Technology, Vol. 9, No. 4, Article 39, Publication date: March 2010.


Received February 2007; revised March 2009; accepted June 2009
A. ONLINE SURVEY OF IMPORTANT PRIVACY FACTORS

We are developing a browser extension that takes in the privacy policy of a website, and automatically analyzes its (usually long and boring) text. The extension then summarizes the privacy policy for you and shows it using visual icons and colors. For example, if the privacy policy allows the website to collect your email address and sell it to third parties, the extension displays an email icon in red or with a danger sign. In order to make a good and useful extension, we need 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. For the purpose of trimming down the extension, 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 it is most important to keep private.

Answer the questions on a scale of 1 to 4, with 1 being “do not care” and 4 being “care a great deal”.

A.1. 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
— Email 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, if you are under 13 years old

A.2. 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

A.3. 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

A.4. Usage
Do you want the extension to inform you 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

A.5. Others
Do you want the extension to inform you about the website's policy for...

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