Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?

Shomir Wilson[1], Florian Schaub[1], Rohan Ramanath[1],
Norman Sadeh[1], Fei Liu[2], Noah A. Smith[3], Frederick Liu[1]

[1]School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
{ shomir, fschaub, rrohan, sadeh, fliu1 }@cs.cmu.edu

[2]Department of Electrical Engineering and Computer Science
University of Central Florida
Orlando, FL, USA
feiliu@cs.ucf.edu

[3]Department of Computer Science and Engineering
University of Washington
Seattle, WA, USA
nasmith@cs.washington.edu

ABSTRACT

Website privacy policies are often long and difficult to understand. While research shows that Internet users care about their privacy, they do not have time to understand the policies of every website they visit, and most users hardly ever read privacy policies. Several recent efforts aim to crowdsource the interpretation of privacy policies and use the resulting annotations to build more effective user interfaces that provide users with salient policy summaries. However, very little attention has been devoted to studying the accuracy and scalability of crowdsourced privacy policy annotations, the types of questions crowdworkers can effectively answer, and the ways in which their productivity can be enhanced. Prior research indicates that most Internet users often have great difficulty understanding privacy policies, suggesting limits to the effectiveness of crowdsourcing approaches. In this paper, we assess the viability of crowdsourcing privacy policy annotations. Our results suggest that, if carefully deployed, crowdsourcing can indeed result in the generation of non-trivial annotations and can also help identify elements of ambiguity in policies. We further introduce and evaluate a method to improve the annotation process by predicting and highlighting paragraphs relevant to specific data practices.

Keywords

Privacy; privacy policies; crowdsourcing; machine learning; HCI.

1. INTRODUCTION

Privacy policies are verbose, often complicated legal documents that provide notices about the data practices of websites and online service providers. McDonald and Cranor [18] showed that if users were to read the privacy policies of every website they access, they would spend an unreasonable fraction of their time doing so; additionally, they found that study participants were largely unable to answer basic questions about what these privacy policies say. Unsurprisingly, many people do not read website privacy policies [10], which are often drafted to ensure legal and regulatory compliance rather than to effectively inform users [28]. Despite these limitations, website privacy policies remain Internet users' primary sources of information on how companies collect, use, and share their data.

Efforts to codify privacy policies, such as the development of the Platform for Privacy Preferences (P3P) standard or more recent initiatives like "Do Not Track" (DNT), have been met with resistance from website operators [4, 8, 17]. While the vast majority of prominent websites have natural language privacy policies (some required by legal regulation [21]), many service providers are reluctant to adopt machine-implementable solutions that would force them to further clarify their privacy practices or to commit to more stringent practices.

In response to this issue, recent efforts have focused on the development of approaches that rely on crowdsourcing to annotate important elements of privacy policies. This includes PrivacyChoice (acquired by AVG), ToS;DR [30], Zimmeck and Bellovin [32], and the Usable Privacy Policy Project [27]. Crowdsourcing is typically applied to tasks that are still difficult for computers to solve, but can be easily solved by humans [22]. Crowdsourcing the extraction of data practices from privacy policies faces a particular challenge: it has been shown that the length and complexity of privacy policies makes them difficult to understand and interpret by most Internet users [11, 18]. Even experts and trained analysts may disagree on their interpretation [25].

In this work, we investigate the feasibility of crowdsourcing privacy policy annotations and explore how the efficiency of crowdworkers can be enhanced by predicting and highlighting policy paragraphs that are relevant to specific data practices. We make the following major contributions:

First, we investigate the accuracy of crowdsourcing the extraction of data practices from privacy policies by comparing crowdworker annotations with those of skilled annotators on a set of 26 privacy policies, focusing on a set of nine annotation questions. By requiring a high level of agreement within a group of crowdworkers ($\geq$80%), we achieve high annotation accuracy (>95%). Crowdworkers generally either match the skilled annotators' interpretation or fail to reach the required agreement level. We find that it is exceedingly rare for crowdworkers to agree on an interpretation that differs from skilled annotators. In other words, while not all annotations are equally easy to crowdsource, a meaningful

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883035.

number of them are amenable to crowdsourcing with a high level of confidence.

Second, we introduce a technique combining machine learning and natural language processing to highlight a small number of paragraphs in a given privacy policy that are likely to be most relevant to a specific annotation task. The results of our respective user study suggest that highlighting paragraphs increases annotation accuracy and that the highlights are perceived as useful by crowdworkers.

The remainder of the paper is organized as follows. We first discuss related work in Section 2. In Section 3, we describe our study design, including our framework for crowdsourcing privacy policy annotations. In Section 4, we analyze the quality of crowdsourced annotations and compare them to the results of skilled annotators. In Section 5, to further improve annotation performance, we propose an approach to automatically identify and highlight paragraphs in a privacy policy that are relevant to a specific annotation question. The effectiveness of this approach has been evaluated in a between subjects study. We describe the study design and results in Section 6. In Section 7, we discuss the implications and benefits of our results, as well as directions for further improving the crowdsourcing of privacy policy annotations.

2. RELATED WORK

The readability issues of privacy policies have been studied extensively [11]. Privacy policies have been evaluated with different readability metrics in different domains, such as energy companies' terms and conditions [16], online social networks [19], and health care notices [9]. Findings suggest that understanding privacy policies requires reading skills and patience that exceed those of the average Internet user.

Multiple efforts have considered extracting data practices from privacy policies with crowdsourcing. For instance, ToS;DR (Terms of Service; Didn't Read) [30] is a community-driven effort to analyze websites' privacy policies and grade their respect for users' privacy. However, ToS;DR's organic and flexible assessment and annotation approach is difficult to scale. Since its inception in 2012, the project has fully or partially analyzed fewer than 70 policies. Zimmeck & Bellovin [32] complement ToS;DR data with automated policy analysis based on natural language processing. Their analysis is limited to a small number of binary questions for which answers are extracted from privacy policies with varying accuracy. Costante et al. [7] use text classification to estimate a policy's completeness based on topic coverage. Other approaches have applied topic modeling to privacy policies [6, 29] and have automatically grouped related sections and paragraphs of privacy policies [15, 24]. Since the complexity and vagueness of privacy policy language makes it difficult to automatically extract complex data practices from privacy policies, we propose to use relevance models to select paragraphs that pertain to a specific data practice and to highlight those paragraphs for annotators.

A common approach to crowdsourcing is to split a complex task into smaller subtasks that are easier to solve [5, 13, 20]. This approach works well for labeling tasks, such as tagging or categorizing images, but privacy policies are substantially more complex: they are lengthy documents filled with legal jargon and often intentional vagueness. Descriptions of a particular data practice may be distributed throughout a policy. For example, in one section a policy may claim that data is not shared with third parties, and later it may list exceptional third parties that receive data. This complexity makes it difficult to correctly interpret a policy's meaning without reading it in its entirety. Thus, a policy's text cannot be trivially partitioned into smaller reading tasks for crowdworkers to annotate in parallel, since integrating contradictory annotations becomes a difficult problem.

Few efforts have been made to crowdsource tasks as complex as annotating privacy policies. André et al. [2] investigate crowdsourcing of information extraction from complex, high-dimensional, and ill-structured data. However, their focus is on classification via clustering, rather than on human interpretation to answer questions. Breaux and Schaub [3] take a bottom-up approach to annotating privacy policies by asking crowdworkers to highlight specific action words and information types in a privacy policy. However, the remaining challenge in their approach is to reconcile results from multiple questions and segments of policy text into a coherent representation of a website's data practices.

The accuracy of policy annotations obtained from crowdworkers has received little prior attention. Reidenberg et al. [25] studied how experts, trained analysts, and crowdworkers disagree when interpreting privacy policies. They conducted a qualitative analysis based on six privacy policies and found that even experts are subject to notable disagreements. Moreover, data practices involving sharing with third parties appeared to be a particular source of disagreement among the annotation groups. On the other hand, Breaux and Schaub [3] found that crowdworkers working in parallel identified more keywords than expert annotators. Both studies were based on a small number of privacy policies (six and five, respectively). In contrast, we assess crowdworkers' accuracy and agreement with trained analysts using a larger set of privacy policies.

3. STUDY DESIGN

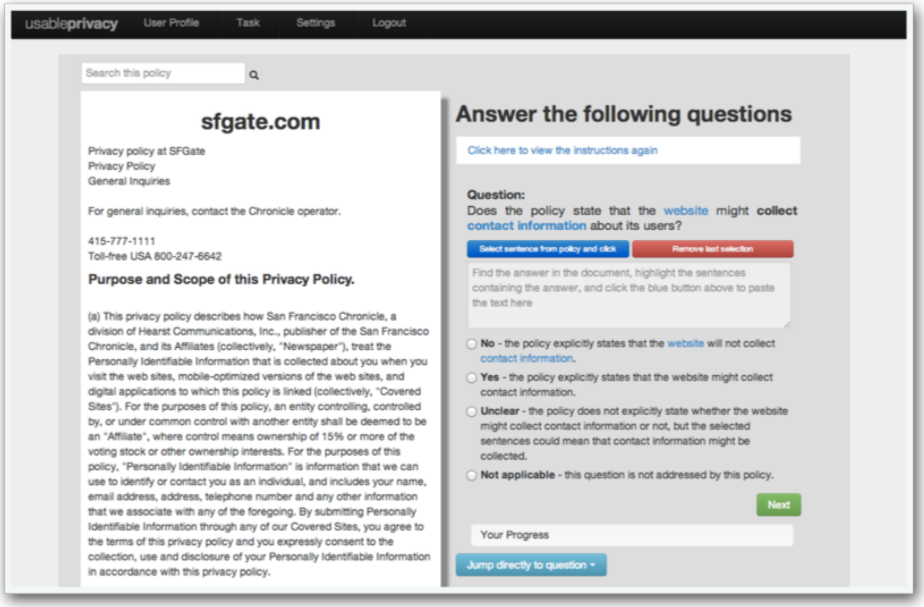
We developed an annotation tool to enable crowdworkers and skilled annotators to annotate privacy policies online. In this section we describe the online annotation tool, our annotation scheme, the privacy policies used in this study, and the two participant groups, namely, skilled annotators and crowdworkers. Carnegie Mellon University's institutional review board approved our study.

3.1 Privacy Policy Annotation Tool

We developed an online annotation tool for privacy policies in order to provide our annotators with an effective interface and workflow to read a privacy policy and answer annotation questions about specific data practices described in the policy. The annotation tool was developed in an iterative user-centered design process that included multiple pilot studies and interview sessions.

The annotation tool, shown in Figure 1, displays a scrollable privacy policy on the left and one annotation question with multiple response options on the right. When selecting an answer, an annotator also selects the passages in the policy text that informed their answer before proceeding to the next question, except when selecting "not applicable." These phrase selections serve as supporting evidence for provided annotations. Multiple text segments can be added to (and removed from) the selection field. The selection field is intentionally located between question and response options to integrate it into the annotator's workflow. Additionally, the annotation tool features a search box above the policy, which enables annotators to search for key terms or phrases within the policy before selecting an answer. While annotators must answer all questions before they can complete a policy annotation task, they can jump between questions, answer them in any order, and edit their responses until they submit the task. This flexibility allows users to account for changes in their interpretation of policy text as they read and understand the privacy policy to answer successive questions.

The policy annotation tool provides users with detailed instructions before they start the task. Users are asked to answer the anno-



sfgate.com

Privacy policy at SFGate
Privacy Policy
General Inquiries

For general inquiries, contact the Chronicle operator.

415-777-1111
Toll-free USA 800-247-6642

Purpose and Scope of this Privacy Policy.

(a) This privacy policy describes how San Francisco Chronicle, a division of Hearst Communications, Inc., publisher of the San Francisco Chronicle, and its Affiliates (collectively, "Newspaper"), treat the Personally Identifiable Information that is collected about you when you visit the web sites, mobile-optimized versions of the web sites, and digital applications to which this policy is linked (collectively, "Covered Sites"). For the purposes of this policy, an entity controlling, controlled by, or under common control with another entity shall be deemed to be an "Affiliate", where control means ownership of 15% or more of the voting stock or other ownership interests. For the purposes of this policy, "Personally Identifiable Information" is information that we can use to identify or contact you as an individual, and includes your name, email address, address, telephone number and any other information that we associate with any of the foregoing. By submitting Personally Identifiable Information through any of our Covered Sites, you agree to the terms of this privacy policy and you expressly consent to the collection, use and disclosure of your Personally Identifiable Information in accordance with this privacy policy.

Answer the following questions

Click here to view the instructions again

Question:
Does the policy state that the website might collect contact information about its users?

Find the answer in the document, highlight the sentences containing the answer, and click the blue button above to paste the text here

○ No - the policy explicitly states that the website will not collect contact information.

○ Yes - the policy explicitly states that the website might collect contact information.

○ Unclear - the policy does not explicitly state whether the website might collect contact information or not, but the selected sentences could mean that contact information might be collected.

○ Not applicable - this question is not addressed by this policy.

Figure 1: The privacy policy annotation tool. It displays a privacy policy (*left*) and one of the annotation questions (*right*). Annotators select phrases from the policy text to support their responses and can jump between questions before submitting the completed annotation task.

tation questions for the main website and to ignore statements about mobile applications or other websites. Users are also instructed to ignore statements applying to a limited audience (e.g., Californians, Europeans, or children). As part of the annotation interface, we provide definitions for privacy-specific terms used in the questions and the response options (e.g., third parties, explicit consent, core service, etc.). Those clarifications are provided as pop-ups when the user hovers over a term highlighted in blue (see Figure 1).

The online annotation tool, the instructions, and the wording of the questions and the response options were refined over multiple iterations. We conducted pilot testing with students and crowdworkers. We also conducted pilot annotations and semi-structured interviews with five skilled annotators to gather feedback, assess the tool's usability, and allow the skilled annotators to familiarize themselves with the tool. Because the skilled annotators provided the gold standard data in our main study, exposing them to the annotation interface at this stage did not affect the results. More specifically, we were interested in eliciting their most accurate interpretations of policies rather than evaluating their interaction with the annotation tool. Pilot tests were conducted with a set of privacy policies different from those used in the actual study. The iterative design resulted in substantial usability improvements. For instance, although we started with a much simpler set of instructions, user tests revealed the need for additional instructions to support the users' interpretation process by reducing ambiguity.

3.2 Annotation Scheme & Selected Policies

We based our annotation scheme on a literature analysis. We identified a small number of data practices and information types that prior work identified as primary concerns for users. We focused on data practices most frequently mentioned in federal privacy litigation and FTC enforcement actions [26], namely collection of personal information, sharing of personal information with third parties, and whether websites allow users to delete data collected about them. In addition, we were interested in how clearly these practices were described with respect to particularly sensitive information types [1, 12, 14]: contact information, financial information, current location, and health information.

Based on relevant data practices we devised a set of nine annotation questions: four questions about *data collection* (Q1–Q4, one for each information type above), four questions about *sharing collected information with third parties* (Q5–Q8), and one question about *deletion of user information* (Q9). For collection and sharing, the provided response options allowed users to select whether a given policy explicitly stated that the website engaged in that practice ("Yes"), explicitly stated that it did not engage in that practice ("No"), whether it was "Unclear" if the website engaged in the practice, or if the data practice was "Not applicable" for the given policy. The sharing questions further distinguished sharing for the sole purpose of fulfilling a core service (e.g., payment processing or delivery), for purposes other than core services, or for purposes other than core services but only with explicit consent. The response options for the deletion question were "no removal," "full removal" (no data is retained), "partial removal" (some data may be retained), "unclear," and "not applicable." Users were instructed to ignore statements concerning retention for legal purposes, as our interest was in annotating retention practices that were questionably motivated but not legally obliged. For all nine questions, each response option was accompanied by an explanation to support its understanding. Throughout the questions, the "unclear" option allowed users to indicate when a policy was silent, ambiguous or self-contradictory with regard to a specific data practice. See Appendix A for the full text of the annotation questions and their response options.

For our study, we selected the privacy policies of 26 news and shopping websites, listed in Table 1. They were selected based on traffic rankings from Alexa.com to provide a cross-section of frequently visited websites. All policies were collected in December 2013 or January 2014.

sfgate.com	costco.com	accuweather.com
money.cnn.com	drudgereport.com	chron.com
bloomberg.com	tigerdirect.com	jcpenney.com
examiner.com	*hm.com*	*washingtonpost.com*
nike.com	ticketmaster.com	*wunderground.com*
abcnews.go.com	bodybuilding.com	*overstock.com*
time.com	*lowes.com*	*barnesandnoble.com*
zappos.com	*shutterfly.com*	latimes.com
bhphotovideo.com	*staples.com*	

Table 1: Privacy policies from 26 shopping and news websites were annotated by crowdworkers and skilled annotators to assess crowdworkers' annotation accuracy. The twelve policies in italics were used in the second experiment to evaluate the effectiveness of highlighting relevant paragraphs.

3.3 Participant Groups

We recruited two participant groups for our study: *skilled annotators*, to obtain gold standard interpretations of privacy policies, and *crowdworkers* to evaluate the accuracy and utility of crowdsourcing privacy policy annotations. Both groups used the same online annotation tool.

The skilled annotators were five graduate students with a background in law and public policy, who concentrated on privacy research and were experienced in reading and interpreting privacy policies. Three of them were female and two were male. They were 23 to 35 years old (median age: 24). Each of the five skilled annotator annotated all 26 policies by answering the nine questions, resulting in 1,170 question responses in total.

Crowdworkers were recruited on Amazon Mechanical Turk. Participants were required to be U.S. residents and to have at least a 95% approval rating for 500 completed tasks. Crowdworkers provided demographics information in an exit survey. Of the crowdworkers, 53.7% were male (117) and 45.9% female (100); 1 crowdworker did not provide their gender. They were 18 to 82 years old (median age: 31). The crowdworkers were somewhat less educated than the skilled annotators: 39% had at least a college degree (bachelor's or higher), 14.7% had only a high school degree, and 2.7% did not complete high school. Primary occupations of the crowdworkers were diverse. The most frequently named occupations were administrative support (12.7%); business, management, or financial (12.4%); computer engineering or information technology (10.6%);service industry (10.1%); student (8.7%); and unemployed (7.8%). The vast majority had no legal training (76.6%). Some (11.5%) indicated that their background in another field provided them with some legal experience. 8.3% indicated they were knowledgeable in legal matters but had no formal legal training. Only 2.3% (5) studied law and 1.4% (3) received other legal or paralegal training. Crowdworkers with legal training were not excluded from participation, since they were part of the population sampled.

Crowdworkers were paid US$6 per annotated privacy policy, and each policy was annotated by ten crowdworkers. The average time for task completion was 31 minutes for the skilled annotators[1]and 24 minutes for the crowdworkers. A total of 218 crowdworkers participated in our study and the vast majority (88.5%) annotated only one policy. We screened task submissions and checked whether question responses were accompanied by meaningful text selections. The rate of bogus answers was extremely low, perhaps due to the approval rating requirements and the relatively high pay.

4. ANALYZING ANNOTATION QUALITY

A major objective of our study was to determine to what extent it is possible to reliably crowdsource meaningful privacy policy annotations and more specifically for the annotation scheme introduced in the previous section. To this end we compared the annotations of our crowdworkers with those produced by our skilled annotators on the dataset of 26 privacy policies.

4.1 Overall Accuracy

In Figure 2, we provide a high-level summary of the accuracy of crowdworker annotations as measured on the 26 privacy policies. In our analysis, we grouped "unclear" and "not addressed in the policy" annotations, since crowdworkers struggled to differentiate between these two options. To consolidate the five skilled annotators' responses, we held them to an 80% *agreement threshold*: for each policy-question pair, if at least four of the five skilled annotators agreed on an answer we considered it to be sufficiently confident for the evaluation standard. Otherwise it was excluded from the comparison. We show results from consolidating crowdworkers' answers using agreement thresholds ranging from 60% to 100% at 10% intervals. Unsurprisingly, higher agreement thresholds yield progressively fewer answers. All crowdworker agreement thresholds demonstrate strong accuracy when evaluated against skilled annotators' answers, with accuracies ranging from 87% (i.e., 132/151 at the 70% crowdworker agreement threshold) up to 98% (42/43 at the 100% crowdworker agreement threshold).

The 80% crowdworker agreement threshold (with 96% accuracy) seems to provide a reasonable balance between accuracy and coverage over the annotations available for analysis. We reached similar conclusions about the skilled annotator agreement threshold, and for the results in the remainder of this paper both agreement thresholds are set at 80%. This suggests that crowdsourcing produces meaningful privacy policy annotations, which match the skilled annotators' interpretation with high accuracy if sufficiently high agreement thresholds are required.

However, the fact that crowdworkers reach that agreement threshold and match the skilled annotators' interpretation for a large fraction of policy-question pairs should not be seen as an indication that privacy policies are clear. Instead, this reflects the fact that annotators were offered answer options that included "unclear" and "not addressed in the policy." For a number of policy-question pairs, skilled annotators and crowdworkers simply agreed with a high level of confidence that the policy was indeed unclear or that an issue was simply not addressed in the policy. Next, we take a detailed look at statistics collected for each of the nine questions. Intuitively, some questions seem to be harder to answer than others.

4.2 Question-Specific Results

Table 2 and Figure 3 provide a detailed comparison of answers from our skilled annotators and our crowdworkers, with both held to 80% agreement thresholds. Some questions appear to be substantially easier to answer than others; for example, our skilled annotators and the crowdworkers found it easy to answer questions about the collection of contact information. However, answering questions about the sharing of contact information seems to be particularly difficult for crowdworkers, who fail to meet the agreement threshold on 23 out of the 26 policies. It is worth noting that some questions seem to be challenging for skilled annotators as well. In particular, skilled annotators fail to converge on 19 of the 26 policy-question pairs dealing with the sharing of financial informa-

[1] This average excludes six assignments with outlier durations greater than 10 hours, where we assume that the skilled annotators stepped away from the task for an extended period of time.

	Skilled Annotators			Crowdworkers		
Question	Yes	Unclear	No Conv.	Yes	Unclear	No Conv.
Collection Contact Info.	26			25		1
Collection Financial Info.	21	4	1	13	4	9
Collection Location Info.	10	12	4	14		12
Collection Health Info.	1	25		1	25	
Sharing Contact Info.	9		17	3		23
Sharing Financial Info.	3	4	19		6	20
Sharing Location Info.	1	20	5		4	22
Sharing Health Info.		25	1		24	2
Deletion of Info.	6	13	7	4	8	14
Total	77	103	54	60	71	103

Table 2: Distributions of skilled annotations and crowdsourced annotations collected for all nine questions across all 26 policies, calculated with an 80% agreement threshold for both groups of annotators. "No Conv." indicates a lack of sufficient agreement among the skilled annotators or crowdworkers. "Yes" indicates that the policy does allow this practice. "Unclear" indicates a "policy is unclear" annotation.

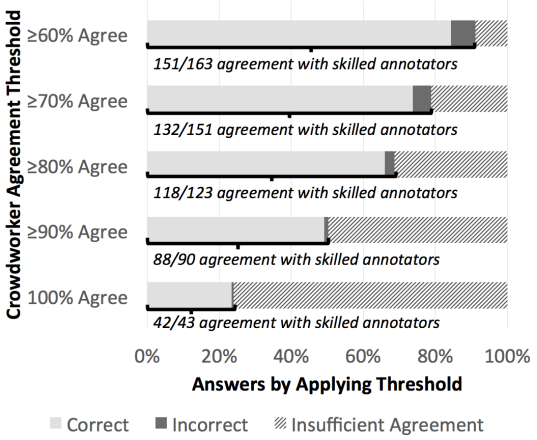


Figure 2: Accuracy of annotations produced by 10 crowdworkers, as measured against skilled annotators, on a set of 179 policy-question pairs. Skilled annotators' answers were held to an 80% agreement threshold (i.e., at least 4 of 5 skilled annotators must agree on the same answer to each policy-question pair to merit its inclusion in the comparison). The bars show crowdworkers' answers when held to a series of progressively higher agreement thresholds.

tion. Overall, we observe that crowdworkers are able to converge on annotations in a majority of cases.

5. HIGHLIGHTING PARAGRAPHS

Our results show that crowdworkers can provide highly accurate privacy policy annotations for some questions, primarily concerning collection and deletion, but that they struggle with questions pertaining to sharing practices, which are typically more spread out in the policy. An exacerbating factor is the length of privacy policies. Policies in our dataset contained 40.8 paragraphs on average, with a standard deviation of 15.8 paragraphs. To fully capture all aspects relating to an annotation question, crowdworkers must read or at least skim the entire policy. This is both time-consuming and sub-optimally efficient, since they must read or skim many paragraphs multiple times as they answer multiple questions. Due to the length of policies, navigating them can be unwieldy, which bears the risk of missing relevant passages in the process.

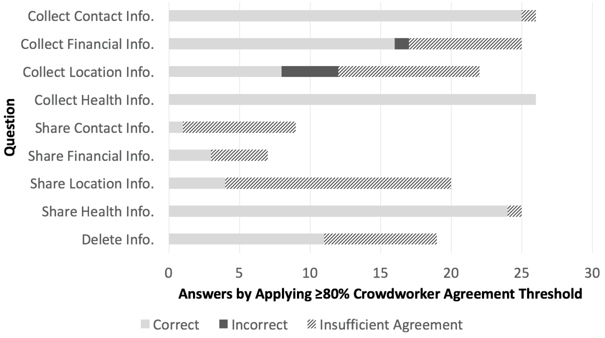


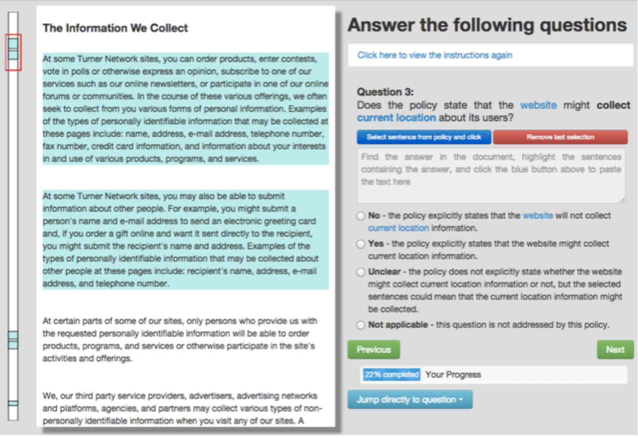
Figure 3: Crowdworkers' annotation accuracy broken down by question. For the sake of this comparison, crowdworkers' answers and skilled annotators' answers were held to 80% agreement thresholds within their cohorts.

As noted before, splitting a policy into smaller parts could reduce reading time, but it bears the risk of losing context and the required holistic view on data practices. Instead, we propose a technique to identify and highlight paragraphs in the policy that are likely to be relevant to the given annotation question. A study evaluating the effects of highlighting paragraphs on annotation accuracy follows in Section 6.

5.1 Identifying Relevant Paragraphs

Our method predicts the top *k* paragraphs in a policy relevant to answering a given question. These relevant paragraphs are highlighted in the annotation tool, as shown in Figure 4, to provide annotators cues about which parts of the policy they should focus on.

We created a separate classifier for each question and applied it to predict each paragraph's relevance to the question. Our approach involves developing regular expressions for a given data practice, which are then applied to a policy's paragraphs. The test selections provided by the skilled annotators were analyzed by a group of five law and privacy graduate students, who picked out phrases (4-10 words) that captured the essence of the response to a specific data practice question. For example, one phrase they chose was "*we obtain . . . information we need to*" (the ellipsis being a placeholder for one or more words). These phrases were first normalized (for



The Information We Collect

At some Turner Network sites, you can order products, enter contests, vote in polls or otherwise express an opinion, subscribe to one of our services such as our online newsletters, or participate in one of our online forums or communities. In the course of these various offerings, we often seek to collect from you various forms of personal information. Examples of the types of personally identifiable information that may be collected at these pages include: name, address, e-mail address, telephone number, fax number, credit card information, and information about your interests in and use of various products, programs, and services.

At some Turner Network sites, you may also be able to submit information about other people. For example, you might submit a person's name and e-mail address to send an electronic greeting card and, if you order a gift online and want it sent directly to the recipient, you might submit the recipient's name and address. Examples of the types of personally identifiable information that may be collected about other people at these pages include: recipient's name, address, e-mail address, and telephone number.

At certain parts of some of our sites, only persons who provide us with the requested personally identifiable information will be able to order products, programs, and services or otherwise participate in the site's activities and offerings.

We, our third party service providers, advertisers, advertising networks and platforms, agencies, and partners may collect various types of non-personally identifiable information when you visit any of our sites. A

Answer the following questions

Click here to view the instructions again

Question 3:
Does the policy state that the website might **collect current location** about its users?

Find the answer in the document, highlight the sentences containing the answer, and click the blue button above to paste the text here

○ **No** - the policy explicitly states that the website will not collect current location information.

○ **Yes** - the policy explicitly states that the website might collect current location information.

○ **Unclear** - the policy does not explicitly state whether the website might collect current location information or not, but the selected sentences could mean that the current location information might be collected.

○ **Not applicable** - this question is not addressed by this policy.

22% completed Your Progress

Select sentence from policy and click
Remove last selection
Previous
Next
Jump directly to question

Figure 4: Privacy policy annotation tool with paragraph highlighting. The five paragraphs most relevant to the shown question are highlighted, and an overview bar (*left*) supports navigation between them. Rather than highlighting only the matched key phrases in the policy, we highlight entire paragraphs to reduce visual clutter and to encourage crowdworkers to read relevant context and thus gain a better understanding of the respective data practice.

stemming and capitalization) and then converted into a list of 110 regular expressions, such as:

```
(place|view|use)(.*?)(tool to collect)(\w+){,3}(inform)
```

In this example, a word with the normalized form *place*, *view*, or *use* must occur in the same sentence as *tool to collect*, and a word with normalized form *inform* (e.g., *information*) must occur within three words of *collect*.

If a regular expression matched one or more paragraphs, those paragraphs were extracted for further feature engineering. After removing stopwords and stemming the selected paragraphs, we used normalized tf-idf values of lower order *n*-grams as features. Thus, for a paragraph, our feature set was comprised of two types of features: (1) *regex features*, i.e., a binary feature for every regular expression in the above constructed list; and (2) *n-gram features*, i.e., tf-idf values for uni-, bi- and trigrams from the extracted paragraphs.

Based on the sentences selected by skilled annotators, we used the respective paragraphs as labels in supervised training. We trained nine classifiers – one for each question – using logistic regression. These classifiers predicted the probability that a given paragraph was relevant to the question for which it is trained. Logistic regression is a standard approach for combining a set of features that might correlate with each other to predict categorical variables. Additionally, it performs well with a low number of dimensions and when the predictors do not suffice to give more than a probabilistic estimate of the response.

Since we were working with a relatively small dataset, we used L_1 regularization to prevent the model from overfitting the data. We used five-fold cross-validation to select the regularization constant. If there are N paragraphs in the corpus, for each of the nine questions, we represent the i^{th} paragraph in the corpus as a feature vector (x_i). Depending on whether it was selected by the skilled annotator or not, we set the label (y_i) as 1 or 0, respectively. The parameters (θ) are learned by maximizing the regularized log likelihood:

$$l(\theta) = \sum_{i=1}^{N} y_i \log h_\theta(x_i) + (1-y_i)\log(1-h_\theta(x_i)) - \lambda\|\theta\|_1.$$

We then pick the top 5 or top 10 paragraphs ordered by probability to constitute the TOP05 and TOP10 relevant paragraph sets for a given policy-question pair.

5.2 Model Validation

To ensure that our model was indeed selecting relevant paragraphs, we calculated the recall of the TOP05 and TOP10 models against the paragraphs selected by the skilled annotators. Across all questions, the average recall rate was 0.91 with a standard deviation of 0.70 for TOP05, and it increased to 0.94 (standard deviation .07) for TOP10. We chose recall as an internal evaluation metric because our goal was to ensure that most of the relevant paragraphs for a question-policy pair were included in the highlights. Highlighting too few paragraphs may have decreased annotation quality, as crowdworkers may have ignored them. Thus, we prefered to potentially highlight some non-relevant paragraphs rather than omitting relevant ones.

6. STUDY: EFFECTS OF HIGHLIGHTING

We integrated the relevance model into our privacy policy annotation tool by color-highlighting the top *k*-relevant paragraphs in each policy, as shown in Figure 4. We also added an overview bar to indicate which parts of the policy were highlighted. Annotators could click on the bar to directly jump to highlighted paragraphs or use buttons above the policy to navigate between highlighted paragraphs. We then conducted a between-subjects study on Mechanical Turk to evaluate the effects of highlighting on annotation accuracy as productivity. We found that highlighting relevant paragraphs can reduce task completion time and positively affect perceived task difficulty without impacting annotation accuracy. Below we describe our study design and results in detail.

6.1 Study Design

Our between-subjects study consisted of a control condition and two treatment conditions that highlighted different numbers of paragraphs (five and ten), in order to investigate the effects of the number of highlights on annotation accuracy and productivity. We named these conditions as follows:

NOHIGH. This control condition consisted of the crowdworkers' responses for the 12 selected policies in the original privacy policy annotation task (cf. Figure 1). Crowdworkers were shown a privacy policy and asked to complete the nine annotation questions. No parts of the policy were highlighted.

TOP05. This condition was identical to NOHIGH, except that for each annotation question the five most relevant paragraphs were highlighted (cf. Figure 4), based on our relevance model.

TOP10. This condition was identical to TOP05, except that the 10 paragraphs most relevant to the shown question were highlighted.

Crowdworkers were recruited on Mechanical Turk and randomly assigned to one of the treatments. If they had participated in the control, they were excluded from further participation, and we ensured that crowdworkers could not participate in more than one condition. In each condition, participants completed the privacy policy annotation task and a short exit survey that gathered user experience feedback and demographic information. We further asked

	Gender		Age		Education	
	Male	Female	Range	Median	High Sch. / Some Coll.	College degree
NOHIGH	53.9%	46.1%	18–82	31	46.7%	41.7%
TOP10	57.5%	42.5%	19–68	29	42.5%	49.1%
TOP05	58.3%	41.7%	20–65	29	46.2%	47.4%

Table 3: Demographics of participants in the highlighting study.

participants to complete an English proficiency test in which they had to fill in missing words in a short passage [31, p. 14]. Each participant annotated only one privacy policy, and we required 10 crowdworkers to annotate a given privacy policy. Participants were compensated with $6 USD. They were required to be US residents with at least a 95% approval rating on 500 HITs. This study received IRB approval.

In order to balance overall annotation costs and annotation scale, we ran the study for a subset of 12 privacy policies randomly selected in equal parts from news and shopping websites. The 12 policies used in the highlighting study are marked in *italics* in Table 1. In total, we obtained annotations from 360 participants.

6.2 Results

We first discuss participant demographics followed by an analysis of the conditions' effect on productivity, accuracy and usability.

6.2.1 *Demographics*

Table 3 summarizes basic demographics for the three participant groups. The three groups exhibited similar characteristics in terms of gender, age, and education level.

Participants reported diverse occupations across all groups. Only 3.6% (NOHIGH), 1.6% (TOP10), and 5% (TOP05) of the crowdworkers reported to work in a position that required legal expertise. However, Figure 5 shows that there is a difference in terms of self-reported legal training between groups. A quarter of the participants (26%) in the NOHIGH group had studied law or received other legal training compared to 3% in the TOP05 and TOP10 groups. This may have been because the NOHIGH annotations were collected at a different time. We carefully considered this aspect in our analysis but did not find it reflected in our accuracy or productivity results. For instance, when asked after the annotation task "*How easy or difficult is it for you to understand legal texts?*" The vast majority in the NOHIGH group rated it as difficult or very difficult (97%), whereas ratings in the TOP05 and TOP10 groups were normally distributed and centered on neutral, as shown in Figure 6. This indicates that the highlighted paragraphs supported the participants' ability to understand the presented legal texts. Additionally, the fraction of correct answers in the English proficiency test were 0.55 (NOHIGH, *SD*=.23), 0.56 (TOP10, *SD*=.24) and 0.55 (TOP05, *SD*=.23), suggesting that English proficiency was comparable across groups.

6.2.2 *Annotation accuracy*

A major concern with drawing annotators' attention to a subset of highlighted paragraphs is that it may negatively impact annotation accuracy, as annotators may miss relevant details in other parts of the policy due to over-reliance on the provided highlights. We evaluated the annotation accuracy of crowdworkers ($\geq$ 80% agreement threshold) against the data previously collected from skilled annotators, focusing on those policy-question pairs from the 12 policies for which at least four of five skilled annotators agreed on the same interpretation ($\geq$ 80% threshold). This was the case for 90 policy-question pairs.

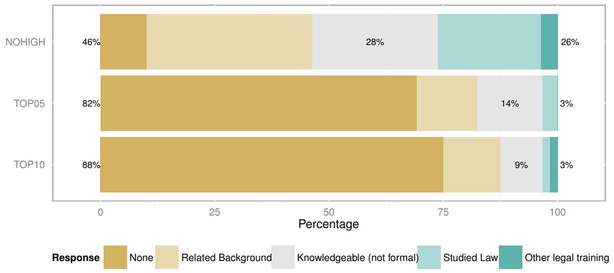


Figure 5: Self-reported level of legal training. Although control group participants (NOHIGH) indicated higher legal training, we observed no measurable effects on annotation accuracy or productivity.

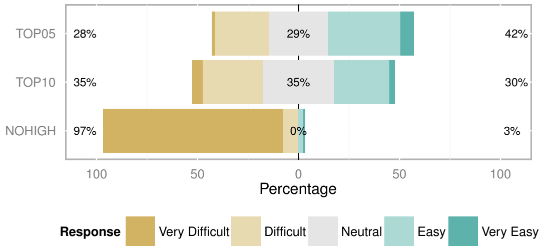


Figure 6: Participants' responses to the question *"How easy or difficult is it for you to understand legal texts?"* Control group participants (NOHIGH) rated their ability substantially lower after completing the annotation task compared to participants in the treatment groups, who were supported by paragraph highlighting.

Figure 7 shows the annotation accuracy for each condition. The annotation accuracy is similar across conditions: 98.1% for NOHIGH and TOP05, and 96.6% for TOP10. This suggests that highlighting relevant paragraphs does not affect annotation accuracy, especially not negatively. In the TOP10 condition, crowdworkers further reached 80% agreement for slightly more policy-question pairs. However, this effect is too small to be directly attributed to the highlighted paragraphs.

We further investigated if highlighting paragraphs affected the crowdworkers' text selections. The goal was to determine whether participants focused solely on the highlighted regions of text, ignoring the rest of the policy, or if they also considered potentially relevant information in non-highlighted parts of the policy. Almost all participants in the treatment conditions self-reported that they either "always read some of the non-highlighted text in addition to the highlighted sections before answering the question" (46.7% TOP05, 46.7% TOP10) or that they "read the non-highlighted text only when [they] did not find the answer within the highlighted text" (53.3% TOP05, 51.7% TOP10). Only 1.6% of participants in the TOP10 group and no one in TOP05 reported that they "never read the non-highlighted text." Additionally, Figure 8 shows the percentage of selections from non-highlighted paragraphs in the policy, for each of the nine annotation questions. For a substantial portion of questions participants selected text from non-highlighted parts of the policy, which confirms that they did not solely focus on the highlights but also considered other policy parts when answering a question. The question-specific variations in Figure 8 sug-

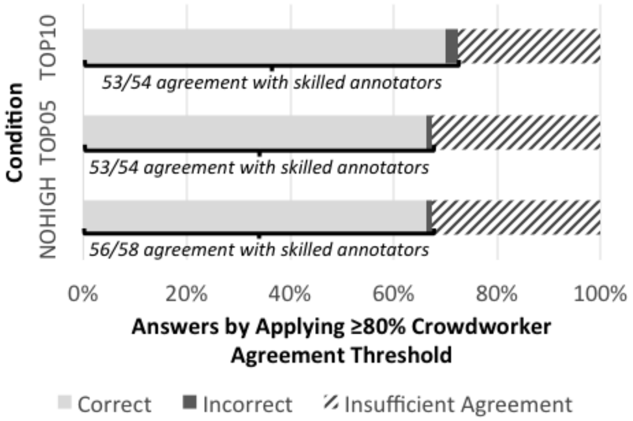


Figure 7: Annotation accuracy in the highlighting study, as measured against skilled annotators ($\geq 80\%$ agreement threshold). Highlighted paragraphs did not negatively affect annotation accuracy.

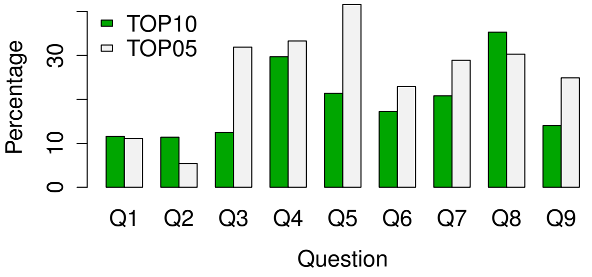


Figure 8: Text selections from non-highlighted parts of a policy for each of the 9 questions. Participants still consider other parts of the policy and do not only focus on highlighted paragraphs.

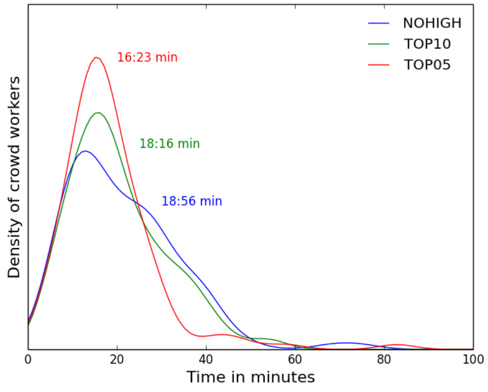


Figure 9: Task completion time in the highlighting study. Highlighting the 5 most relevant paragraphs substantially reduces median task completion time.

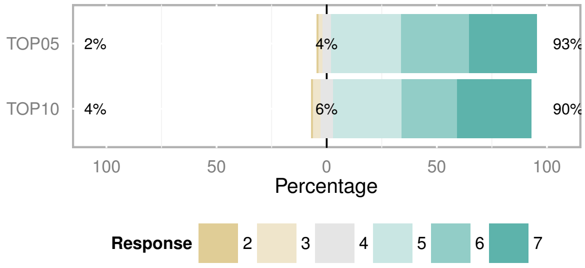


Figure 10: Perceived usefulness of highlighting paragraphs in the treatment conditions (from (1) not at all helpful to (7) very helpful).

gest that some questions may benefit from the use of different machine learning methods, but highlighting relevant paragraphs does not seem to bias annotators to ignore non-highlighted parts of the policy.

However, while both groups selected text from non-highlighted parts for all questions of the policy, TOP05 participants tended to select more information from non-highlighted parts. This suggests that, for some questions, more than 5 paragraphs need to be considered to fully grasp a data practice. We can further observe differences for certain annotation questions and data practices. For instance, collection of financial (Q3) and health information (Q4) practices are often not as explicitly and concisely addressed as collection of contact (Q1) or location (Q2) information.

6.2.3 Productivity & Usability

We further analyzed how highlighting paragraphs affected the crowdworkers' productivity in terms of task completion time, as shown in Figure 9. The median task completion times for the three conditions were 18 min 56 sec (NOHIGH), 18 min 16 sec (TOP10), and 16 min 23 sec (TOP05). Although these differences were not statistically significant (Kruskal-Wallis test), we observe that highlighting five paragraphs appeared to substantially reduce task completion time by more than 2 minutes without impacting annotation accuracy. Highlighting 10 paragraphs had only a marginal effect on task completion time, suggesting that crowdworkers in the control condition may have read or skimmed a similar number of paragraphs.

We further asked participants in the TOP05 and TOP10 groups to rate the perceived usefulness of paragraph highlighting on a seven-point scale ranging from *"Not at all helpful"* (1) to *"Very Helpful"* (7). Distribution of answer choices are shown in Figure 10. The median answer choice was Helpful (6) for both groups, signifying that the highlighted paragraphs were seen as useful cues and likely supported the annotators in determining the answer for a given data practice question. Additionally, as shown in Figure 6, the control group (NOHIGH) rated their ability to understand legal text substantially lower after completing the annotation task compared to participants in the treatment conditions – despite more legal training in this group (cf. Figure 5).

Thus, we infer that paragraph highlighting in the annotation tool improved annotation productivity and user experience, which is an important factor for worker retention and cultivating a crowd of experienced annotators. Simultaneously, paragraph highlighting did not negatively impact annotation accuracy.

7. DISCUSSION AND CONCLUSIONS

The results presented in our studies show promise for using crowdworkers to answer questions about privacy policies. It appears that data practices can be reliably extracted from privacy policies through crowdsourcing. In other words, crowdsourcing is potentially a viable process to provide the data required for new types of browser plug-ins and other users interfaces aimed at informing Internet users, who have generally given up on trying to read privacy policies. Furthermore, crowdsourcing could aid privacy policy analysis and ease the work of regulators, who currently rely on manual inspection by experts for policy sweeps.

Our results further show that crowdsourcing privacy policy annotations is not trivial. We went through multiple iterations to refine our task design, as well as the annotation questions and response options. Given the vagueness of privacy policies it was essential to provide crowdworkers with annotation options that indicate that a policy is unclear or does not address a given issue. Considering that even privacy experts may not always agree on the interpretation of policies [25], we cannot expect crowdworkers to perform better. From a public policy standpoint, these annotation options could also help identify egregious levels of ambiguity in privacy policies, either in response to particular types of questions or at the level of entire sectors. Finally, policy-question pairs where crowdworkers cannot converge could also be the basis for processes that engage website operators to clarify their practices.

Although the 80% crowd agreement threshold appears promising, additional experiments with a larger number of policies will need to be conducted to further validate our results. An opportunity also exists for a user study to understand how to meet users' needs more precisely. Additional opportunities for refining this line of inquiry include allowing crowdworkers to rate the difficulty of answering a specific annotation question for a given policy. These ratings could then be considered in the aggregation of results. Such ratings, as well as the performance of individual crowdworkers, could also be used to develop more versatile crowdsourcing frameworks, where different crowdworkers are directed to different annotation tasks based on their prior performance and where the number of crowdworkers is dynamically adjusted. The longitudinal performance of crowdworkers could be monitored in order to place more weight on high-performing workers. These and similar approaches [23] could be used to dynamically determine and allocate the number of annotations required for a question-policy pair. Additionally, the use of skilled workers on freelancing platforms such as Upwork and NC may reduce the amount of redundancy necessary to reach answers with confidence.

Our research also shows that techniques that highlight paragraphs relevant to specific annotation questions can help increase productivity and improve the user experience, as workers are provided with cues about which paragraphs they should focus on. This is important given the length of privacy policies and the way the discussion of some data practices is often spread across the text of policies. The number of highlighted paragraphs plays an essential role. In our study, highlighting the five most relevant paragraphs decreased task completion time, but also resulted in more text being selected from non-highlighted areas compared to highlighting 10 paragraphs. Ideally, we would want to highlight just enough for the annotator to clearly identify the answer. Thus, we are currently investigating approaches to dynamically adapt the number of highlights to question-specific parameters. For instance, some data practices such as collection of contact information are plainly stated in one part of the policy, while others require annotators to pay attention to multiple policy parts, such as third party sharing practices. We plan to fine-tune our relevance models in the future and explore extensions of our approach to additional data practices.

We further plan to extend our annotation efforts to more and other categories of websites. While our current sample of 26 privacy policies is not representative of all privacy policies, we are confident that our results provide meaningful indications of the annotation difficulty of collection, sharing, and deletion practices, which may vary for individual websites but are largely ubiquitous.

Our goal is to further improve our annotation framework in order to improve the quality and cost efficiency of privacy policy annotations. We want to enable large-scale privacy policy analysis. Our contributions in this work show the promise of crowdsourcing privacy policy annotations and the potential of achieving the required scale-up by combining crowdsourcing with machine learning and natural language processing to enhance crowdworker productivity.

Acknowledgements

This research has been partially funded by the National Science Foundation under grant agreement CNS-1330596.

APPENDIX

A. ANNOTATION QUESTIONS

Questions Q1 through Q4 address the collection of contact information, financial information, current location information, and health information, respectively. Their wording is largely identical, and for brevity, only Q1 and its answers are shown below.

Q1: Does the policy state that the website might **collect contact information** about its users?

- **No** – the policy explicitly states that the website will not collect contact information.
- **Yes** – the policy explicitly states that the website might collect contact information.
- **Unclear** – the policy does not explicitly state whether the website might collect contact information or not, but the selected sentences could mean that contact information might be collected.
- **Not applicable** – this question is not addressed by this policy.

Questions Q5 through Q8 address the sharing of contact information, financial information, current location information, and health information, respectively. Their wording is largely identical, and for brevity, only Q5 and its answers are shown below.

Q5: Does the policy state that the website might **share contact information** with **third parties**? Please select the option that best describes how contact information is shared with third parties. Please ignore any sharing required by law (e.g., with law enforcement agencies).

- **No sharing** – the policy explicitly states that the website will not share contact information with third parties.
- **Sharing for core service only** – the policy explicitly states that the website might share contact information with third parties, but only for the purpose of providing a core service, either with explicit or implied consent/permission from the user.

- **Sharing for other purpose** – the policy explicitly states that the website might share contact information with third parties for other purposes. The policy makes no statement about the user's consent/permission or user consent is implied.
- **Sharing for other purpose (explicit consent)** – the policy explicitly states that the website might share contact information with third parties for a purpose that is not a core service, but only if the user provided explicit permission/consent to do so.
- **Unclear**
- **Not applicable**

Finally, Q9 addresses deletion of personal data.

Q9: What is the website's policy about letting its users **delete their personal data**? Please ignore any statements concerning retention for legal purposes.

- **No removal** – the policy explicitly states that the user will not be allowed to delete their personal data.
- **Full removal** – the policy explicitly states that users may delete their personal data and that no data will be retained for any purpose, whether the data was provided directly by the user, generated by the user's activities on the website, or acquired from third parties.
- **Partial removal** – the policy explicitly states that users may delete their personal data but some/all of the data might be retained for other purposes, whether the data was provided directly by the user, generated by the user's activities on the website or acquired from third-parties.
- **Unclear**
- **Not applicable**

B. REFERENCES

[1] M. S. Ackerman, L. F. Cranor, and J. Reagle. Privacy in e-commerce: Examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, EC '99, pages 1–8, New York, NY, USA, 1999. ACM. 00456.

[2] P. André, A. Kittur, and S. P. Dow. Crowd synthesis: Extracting categories and clusters from complex data. In *Proc. CSCW '14*, pages 989–998. ACM, 2014.

[3] T. D. Breaux and F. Schaub. Scaling requirements extraction to the crowd. In *RE'14: Proceedings of the 22nd IEEE International Requirements Engineering Conference (RE'14)*, Washington, DC, USA, August 2014. IEEE Society Press.

[4] J. Brookman, S. Harvey, E. Newland, and H. West. Tracking compliance and scope. *W3C Working Draft*, November 2014.

[5] L. B. Chilton, G. Little, D. Edge, D. S. Weld, and J. A. Landay. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1999–2008. ACM, 2013.

[6] P. Chundi and P. M. Subramaniam. An Approach to Analyze Web Privacy Policy Documents. In *KDD Workshop on Data Mining for Social Good*, 2014.

[7] E. Costante, Y. Sun, M. Petković, and J. den Hartog. A machine learning solution to assess privacy policy completeness. In *Proc. of the ACM Workshop on Privacy in the Electronic Society*, 2012.

[8] L. Cranor, B. Dobbs, S. Egelman, G. Hogben, J. Humphrey, M. Langheinrich, M. Marchiori, M. Presler-Marshall, J. Reagle, D. A. Stampley, M. Schunter, and R. Wenning. The Platform for Privacy Preferences 1.1 (P3P1.1) Specification. Working group note, W3C, 2006. http://www.w3.org/TR/P3P11/.

[9] T. Ermakova, B. Fabian, and E. Babina. Readability of Privacy Policies of Healthcare Websites. In *12. Internationale Tagung Wirtschaftsinformatik (Wirtschaftsinformatik 2015)*, 2015.

[10] Federal Trade Commission. Protecting consumer privacy in an era of rapid change: Recommendations for businesses and policymakers, 2012.

[11] C. Jensen and C. Potts. Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proc. CHI '04*. ACM, 2004.

[12] A. N. Joinson, U.-D. Reips, T. Buchanan, and C. B. P. Schofield. Privacy, trust, and self-disclosure online. *Human-Computer Interaction*, 25(1):1–24, Feb. 2010.

[13] A. Kittur, B. Smus, S. Khamkar, and R. E. Kraut. Crowdforge: Crowdsourcing complex work. In *Proc. UIST '11*, pages 43–52. ACM, 2011.

[14] P. G. Leon, B. Ur, Y. Wang, M. Sleeper, R. Balebako, R. Shay, L. Bauer, M. Christodorescu, and L. F. Cranor. What matters to users?: Factors that affect users' willingness to share information with online advertisers. In *Proc, SOUPS '13*. ACM, 2013.

[15] F. Liu, R. Ramanath, N. Sadeh, and N. A. Smith. A step towards usable privacy policy: Automatic alignment of privacy statements. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, 2014.

[16] E. Luger, S. Moran, and T. Rodden. Consent for all: Revealing the hidden complexity of terms and conditions. In *Proc. CHI '13*. ACM, 2013.

[17] A. M. McDonald. Browser Wars: A New Sequel? The technology of privacy, Silicon Flatirons Center, University of Colorado, 2013. presented Jan. 11, 2013.

[18] A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *I/S: J Law & Policy Info. Soc.*, 4(3), 2008.

[19] G. Meiselwitz. Readability Assessment of Policies and Procedures of Social Networking Sites. In *Proc. OCSC '13*, 2013.

[20] M. Negri, L. Bentivogli, Y. Mehdad, D. Giampiccolo, and A. Marchetti. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 670–679. Association for Computational Linguistics, 2011.

[21] Official California Legislative Information. The Online Privacy Protection Act of 2003, 2003.

[22] A. J. Quinn and B. B. Bederson. Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1403–1412, New York, NY, USA, 2011. ACM. 00257.

[23] N. Quoc Viet Hung, N. T. Tam, L. Tran, and K. Aberer. An evaluation of aggregation techniques in crowdsourcing. In *Proc. WISE '13*, pages 1–15. Springer, 2013.

[24] R. Ramanath, F. Liu, N. Sadeh, and N. A. Smith. Unsupervised alignment of privacy policies using hidden Markov models. In *Proc. ACL '14*, 2014.

[25] J. R. Reidenberg, T. D. Breaux, L. F. Cranor, B. French, A. Grannis, J. T. Graves, F. Liu, A. M. McDonald, T. B. Norton, R. Ramanath, N. C. Russell, N. Sadeh, and F. Schaub. Disagreeable privacy policies: Mismatches between meaning and users' understanding. In *Proceedings of 42nd Research Conference on Communication, Information and Internet Policy*, TPRC'14, 2014.

[26] J. R. Reidenberg, N. C. Russell, A. J. Callen, S. Qasir, and T. B. Norton. Privacy harms and the effectiveness of the notice and choice framework. *I/S: Journal of Law & Policy for the Information Society*, 11, 2015.

[27] N. Sadeh, A. Acquisti, T. D. Breaux, L. F. Cranor, A. M. McDonald, J. R. Reidenberg, N. A. Smith, F. Liu, N. C. Russell, F. Schaub, and S. Wilson. The usable privacy policy project: Combining crowdsourcing, machine learning and natural language processing to semi-automatically answer those privacy questions users care about. Tech. report CMU-ISR-13-119, Carnegie Mellon University, 2013.

[28] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 1–17, Ottawa, July 2015. USENIX Association.

[29] J. W. Stamey and R. A. Rossi. Automatically identifying relations in privacy policies. In *Proc. SIGDOC '09*. ACM, 2009.

[30] Tos;DR. Terms of service didn't read, 2012. http://tosdr.org/ (accessed: 2015-03-11).

[31] University of Cambridge. Certificate of proficiency in english (CPE), CEFR level C2): Handbook for teachers, 2013.

[32] S. Zimmeck and S. M. Bellovin. Privee: An architecture for automatically analyzing web privacy policies. In *USENIX Security Symposium*. USENIX, 2014.