Learning Privacy Expectations by Crowdsourcing Contextual Informational Norms

Yan Shvartzshnaider
New York University

Schrasing Tong
Princeton University

Thomas Wies
New York University

Paula Kift
New York University

Helen Nissenbaum
New York University

Lakshminarayanan Subramanian
New York University

Prateek Mittal
Princeton University

Abstract

Designing programmable privacy logic frameworks that correspond to social, ethical, and legal norms has been a fundamentally hard problem. Contextual integrity (CI) (Nissenbaum 2010) offers a model for conceptualizing privacy that is able to bridge technical design with ethical, legal, and policy approaches. While CI is capable of capturing the various components of contextual privacy in theory, it is challenging to discover and formally express these norms in operational terms.

In the following, we propose a crowdsourcing method for the automated discovery of contextual norms. To evaluate the effectiveness and scalability of our approach, we conducted an extensive survey on Amazon's Mechanical Turk (AMT) with more than 450 participants and 1400 questions. The paper has three main takeaways: First, we demonstrate the ability to generate survey questions corresponding to privacy norms within any context. Second, we show that crowdsourcing enables the discovery of norms from these questions with strong majoritarian consensus among users. Finally, we demonstrate how the norms thus discovered can be encoded into a formal logic to automatically verify their consistency.

Introduction

Responding to a widely accepted social need, technical communities in academia as well as in industry have conscientiously worked towards protecting and promoting privacy in their respective fields of expertise. An important research challenge in real world systems, and a key requirement of the privacy-by-design initiative (Computing Community Consortium 2015), is to incorporate meaningful conceptions of privacy; either those that are expressed explicitly in the law, or – equally important – those that implicitly shape ethical and societal expectations.

While law- and policymakers primarily build on existent privacy regulation and legal precedents, in contexts where advances in information technology and digital media have critically affected baseline practices, it makes sense to reevaluate and respond to social norms and expectations. In large part this is because privacy laws-on-the-books and legal precedents reflect a world prior to these advances. A key instance is education, where the Family Educational Rights and Privacy Act (FERPA), enacted in 1974, long predates the rise of heavily used social platforms such as Facebook, Google, not to mention a slew of third-party education services and digital learning platforms, including MOOCs that have radically affected flows of personal information. The privacy expectations of ordinary users are likely to evolve hand-in-hand with new technologies and may no longer adequately be reflected by laws such as FERPA. Thus while the efforts to formalize explicit laws remain important, there is also a need to discover and formally express evolving societal norms. This is the challenge that our work addresses, namely, the discovery and formal expression of social norms according to the framework of contextual integrity.

Contextual integrity (CI) (Nissenbaum 2010) offers a model for conceptualizing privacy that is able to bridge technical design with ethical, legal, and policy approaches. It postulates that privacy is neither about secrecy nor control but about the appropriate flow of information within a particular context, where appropriateness is defined as compliance with informational norms. CI posits a five-element tuple to distill contextual informational norms that reflect where or from whom the information flow originates *(sender)*, what type of information is being conveyed *(attribute)*, about whom *(subject)* and to whom *(recipient)*. CI thus offers a common language that is able to express both privacy policies and expectations using a single structure.

Past work that has taken norms, or rules, as a given, e.g. taking legal rules as points of departure, has largely focused on developing formal languages and logical frameworks for expressing these, detecting infractions, and developing approaches for accountability and enforcement (Chowdhury, Gampe, and others 2013; Barth, Datta, and others 2006; Criado and Such 2015). While the discovery of norms is usually beyond the scope of their work, they have made significant contributions to the technical field, generating machine-readable access rules and implementing complex constraints that map given rules.

In our work, we propose a crowdsourcing method for the automated discovery of contextual norms. Given a specific context, our crowdsourcing methodology automatically generates several context-specific privacy rule candidates, expressed in the CI format. These, then, are presented to users in a crowdsourcing platform. Based on users' preferences expressed on Amazon's Mechanical Turk (AMT), we have

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

sought to extract norms that enjoy majoritarian consensus. In principle, such work precedes formal expression by helping to discover contextual information norms that might be neglected by approaches that depend on explicitly state rules.

A system based on the norms that our methodological approach allows us to uncover could then be enhanced with machine learning capacities, and thus allowing it to adjust to the continuing evolution of norms, mirroring the complexities of social life offline. Although our approach to discovering norms and expressing them in a logic constitutes an alternative to those relying on pre-conceived social rules–whether expressed in the law or articulated by experts–it is important to note that we do not reject these alternatives. Instead, we are offering an additional, systematic source that is particularly well suited for a range of systems, including information systems with social actors. Specifically, our paper makes the following contributions:

1. **A crowdsourcing methodology for discovering informational (privacy) norms for a given context.** We elicit informational norms based on a crowdsourcing approach that queries users on their privacy expectations based on automatically generated privacy statements using the language of CI.
2. **Converting crowd-sourced responses to an Effectively Propositional Logic (EPR) form.** We provide a framework for verifying the consistency of consensus-based crowdsourced responses and derive a formal representation of the informational norms in EPR.
3. **Generalization to other contexts.** Our methodology can be adapted and generalized to derive consistent crowdsourced informational norms in other social contexts.

Crowdsourcing Contextual Privacy Norms

People learn and adopt implicit and explicit informational norms through interactions with their families, friends, and communities; by watching how people behave and how they react to other people's behavior; from educational training, the arts, and cultural activities; from the study of law and policy, and so forth. Consequently, capturing relevant operational norms is not an easy and straightforward process. Our personal preferences and our interests may deviate from what society collectively agrees upon; law and regulation, handbooks, as well as expert opinions may lag behind the current state of privacy expectations of the majority. These various sources of privacy rules play different roles in the regulation of privacy in society.

Contextual Integrity (CI)

The theory of Contextual Integrity (CI) argues that privacy is not retained using a single-argument function that accepts as a parameter *what* to hide (Posner 1977) or control (Westin Alan 1967). Rather, it can be viewed as a derivative of an informational norm (function) that reflects the appropriateness of an informational flow between *actors* (stakeholders) in a given context.

Contextual informational norms accept five parameters to define which sender (*sender*) gets to share what type of information (*attribute*) with a particular recipient (*recipient*) about a particular information subject (*subject*) under certain conditions (*transmission principles*). For example, as depicted in Figure 1, in the educational setting, we might expect a professor to share the grades of her students with her teaching assistant with the requirement of confidentiality. In other words, we generally accept an informational flow where a professor, the sender, provides students' grades, the subject and attribute respectively, to her TA, the recipient, with the transmission principle of confidentiality. Any variation in the parameters alters the information flow and hence potentially violates our privacy expectations, for instance, if the TA shares one of the student's grades with others students in the class. This would constitute an alternative informational flow which may or may not correspond to our privacy expectations.

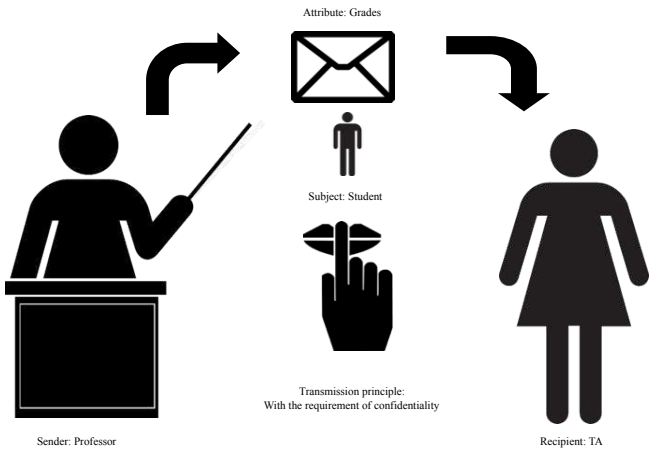

Figure 1: Information norm: The professors is allowed to share the student's grade with the student's TA with the requirement of confidentiality.

While CI, as a conceptual framework, is capable of revealing different aspects of contextual privacy, it does not provide a method for discovering these norms on the ground. Past efforts (Barth, Datta, and others 2006; Criado and Such 2015; Krupa and Vercouter 2012) have, for the most part, taken certain privacy norms for granted, and have focused on deriving logical frameworks to express and enforce them. Automating the discovery of informational norms remains an important open question – one that we address in our work.

Complicating the situation, CI recognizes that norms are constantly in the process of becoming, hence informational norms can also evolve as the sociotechnical environment changes. Although the theory of CI has a prima facie preference for entrenched informational norms, it allows for normative transformations when the resultant norms can better promote the values, goals, and ends of a given context.

Here, we are interested in formally capturing CI's bedrock notion of appropriate information flows, that is, entrenched privacy expectations, formally represented by CI's context-specific information flows. Towards this end, we have de-

veloped a systematic method for discovering these norms, namely by crowdsourcing users' privacy expectations in an online environment. We then analyze users' preferences to derive an implicit consensus on a set of privacy norms that the majority of users will collectively accept.

Crowdsourcing Privacy Norms

To elicit the users' privacy preferences in a particular context we translated norms formulated in the language of CI into corresponding questions and subsequently presented these to crowd-sourced workers. Our examples are based on the educational context which is the context we are most familiar with.

As discussed in the previous section, a CI norm comprises the following elements: roles of actors (senders, subjects and recipients), attributes (information types), and transmission principles (constraints on flow). That is, each norm can be represented as a 5-tuple using these parameters. See below for possible values for the CI parameters in the educational context.

Examples of Roles: Students, Professors, TAs, Registrar, University IT staff, academic advisor

Examples of Attributes: Grades, Transcript, Name, Email address, Address, Record of attendance, Level of participation in class, Photo, Library records, Contents posted on online learning systems (e.g., Blackboard, Classes, etc.), term paper

Examples of Transmission Principles:

- **Knowledge:** If the ⟨ sender ⟩ let the ⟨ subject ⟩ know
- **Permission:** If the ⟨ sender ⟩ asked for the ⟨ subject's ⟩ permission
- **Breach of contract:** If the ⟨ subject ⟩ is performing below a certain standard

After identifying possible values for the roles, attributes, and transmission principles in the context of interest, we inject each of the relevant CI parameters into the following Yes-or-No question template:

"*Is it acceptable for the ⟨sender⟩ to share the ⟨subject⟩'s ⟨attribute⟩ with ⟨recipient⟩⟨transmission principle⟩?*"

Using the resulting questions we design a survey that we submit to a crowdsourcing platform. We then approximate the users' privacy expectations in the given context by analyzing the answers to the crowdsourced survey using the indicators we describe below.

State space reduction. The number of possible questions to be generated grows polynomially with the number of values for the CI parameters. Thus, the space of survey questions can be large, making their exhaustive generation a challenging proposition. For instance, even for the simple classroom context outlined above, exhaustive enumeration yields more than twenty-thousand questions. However, we can exploit the privacy experts' domain knowledge about the specific context to reduce the state space that needs to be considered for enumeration. Often specific attributes only make sense for subjects that have specific roles. For instance, in the classroom example, all the listed attributes only apply to students. In turn, this means that *student* is the only subject that we need to consider in our questions. This restriction alone reduces the size of the state space by a factor of five. Similar constraints can be formulated for feasible combinations of senders, receivers, and attributes. For highly complex contexts, the remaining state space may still be too large, even after eliminating irrelevant information flows. In such cases, the privacy expert can help identify the "regions" of the state space that describe the bulk of the information flows that are observable in practice. The initial survey will then focus on these regions.

Survey design For our purposes, we took the educational context as an example to test whether the CI framework would be able to better capture users' privacy expectations. We constructed a context-specific set of questions that would allow us to crowdsource corresponding informational norms. Our target population was United States residents, between 18-26 years of age, and currently enrolled in (or graduated within the past three years from) an institution of higher education in the U.S. We posed these questions using an online survey designed with Qualtrics and administered on Amazon's Mechanical Turk (AMT).

We used a script to generate the initial set of norms based on the most common CI parameters in a classroom setting (e.g., teachers and students as actors; grades as attribute; knowledge or consent as transmission principles). We then translated these norms into the corresponding questions. As discussed earlier, we do not enumerate the full space of all possible privacy norms that can be expressed over the given parameter values. Instead, we rely on the input of two domain experts to reduce the explored space to those candidate norms that cover the bulk of the relevant information flows. To decrease the total number of questions asked, two domain experts performed a preliminary scan of the norms to identify the ones that clearly did not make any sense. Rather than manually going through the questions one by one, the authors focused exclusively on valid pairs of senders and attributes. Based on the feedback, we introduced constraints to remove irrelevant questions (e.g., university librarians cannot be senders of content posted on online learning systems). Following these restrictions, we ended up with a total of 1411 questions. We randomized the questions and divided them up into 16 sets (13 with 88 questions, 3 with 89 questions) with about 30 respondents each. That way, we would be able to ask all possible questions within the context (i.e., achieve completeness) at a reasonable cost ($2 per user per survey, plus AMT fees).

Furthermore, we provided users with several "Does not make sense" (DMS) options[1] for questions that suggest implausible scenarios. For example, some questions present scenarios that may be structurally sound, but are simply unlikely to occur in a real world system and thus make little sense to the survey participants, e.g. when certain senders

[1] 1) The sender is unlikely to have the information, 2) The receiver would already have the information, 3) The question is ambiguous

are unlikely to have access to certain attributes.

In total, we had 451 respondents to the 16 surveys: each user had to respond to 88-89 questions, with 28-32 respondents per question in each survey. The average completion time of the survey was around 14 minutes per user. We believe the academic community will benefit from the dataset and therefore made it public[2].

Approximation of Users' Privacy Expectations

We introduce a number of indicators that together allow us to develop an estimate of the users' overall attitude towards a preexisting set of privacy norms. Specifically, we considered three metrics in our evaluation: the *norm approval score*, the *user approval score* and the *divergence score*.

Norm approval score (NA) This is our measure of what norm is approved by the community based on the users' answers (scores) to the question corresponding to it. We define the *norm approval score (NA)* of question i as follows:

$$NA_i = \frac{\sum_{j=1}^{m} Y_{i,j}}{\sum_{j=1}^{m} (Y_{i,j} + N_{i,j} + DMS_{i,j})} = \frac{\sum_{j=1}^{m} Y_{i,j}}{m} \quad (1)$$

Here, $Y_{i,j}$ is defined to be 1 iff respondent j answered "Yes" to question i. Similarly, $N_{i,j}$ and $DMS_{i,j}$ indicate whether user j answered "No", respectively, chose "Does not make sense". Thus, NA_i is the ratio between the total number of "Yes" answers[3] and the number of all answers for question i across all m respondents. A norm is considered approved if its NA exceeds a certain threshold, e.g., a simple majority ($> 50\%$) or two-third majority ($> 66\%$)

User approval score (UA) This metric measures the relative number of norms that have been approved by a given respondent. Formally, the value UA_j for respondent j is defined as

$$UA_j = \frac{\sum_{i=1}^{n} Y_{i,j}}{\sum_{i=1}^{n} (Y_{i,j} + N_{i,j} + DMS_{i,j})} = \frac{\sum_{i=1}^{n} Y_{i,j}}{n} \quad (2)$$

where n is the total number of questions in the survey that j responded to.

Divergence score (DS) This metric looks at how the answers of individual respondents vary from the norms that have been approved or disapproved by the whole community, subject to a given NA threshold. Intuitively, it quantifies how dissatisfied a user is with the extracted set of operational norms. Formally, the divergence score DS_j of respondent j is defined as

$$DS_j = \sum_{i=1}^{n} c_i \oplus u_{i,j} \quad (3)$$

Here, the bit $u_{i,j}$ is defined to be 1 iff respondent j approved the norm described by question i and c_i is defined to be 1 iff the community as a whole approved the norm. Hence, DS_j indicates the number of times respondent j's expectations differed from the operational privacy rule set that was enforced based on the chosen NA threshold.

Verification of Extracted Rules

We use formal verification technology to analyze the consistency of the derived privacy logic. The crowd-sourced privacy rules can be encoded in formal logic. Such an encoding enables us to employ formal verification technologies, such as automated theorem provers, for detecting potential logical inconsistencies in the rules. These inconsistencies may suggest hidden underlying assumptions made by the survey participants and problems with the chosen NA threshold. Thus, by using formal verification techniques, we can automate the process of checking our privacy rules for consistency and hence aid the overall survey design.

More specifically, we encode the derived privacy rules into so-called *Effectively Propositional Logic (EPR)*[4]. EPR is a decidable fragment of first-order predicate logic and there exist several automated theorem provers that implement decision procedures for this fragment. We can use these tools to automate the verification tasks of interest. In the following, we describe the basic idea behind this encoding.

Our encoding of CI rules into first-order logic uses specific predicates that model the relevant relationships between the different CI parameters. Central to the encoding is the predicate

$$\mathsf{Allowed}(ctx, sndr, recp, subj, attr)$$

which expresses that in context ctx, actor $sndr$ is allowed to send information on attribute $attr$ of actor $subj$ to actor $recp$. The Allowed predicate thus represents all flows that are admissible in each context. The predicate is given meaning by logical constraints that encode the derived privacy rules. In order to be able to express these rules in predicate logic, we introduce auxiliary context-specific predicates to encode roles and relationships between individual actors as well as transmission principles. For example, in the class room context, we may have the auxiliary predicates $\mathsf{Professor}(a)$ and $\mathsf{ParentOf}(a, b)$, which encode that actor a is in the role of *professor*, respectively that, a is in the role of *b's parent*.

Each individual CI rule is then of the form $R(ctx, sndr, recp, subj, attr, tr)$ where R is a conjunction involving the auxiliary predicates and (dis)equalities over the given flow parameters. Once the context-specific predicates for expressing the rules are fixed, the encoding of survey questions to rules can be easily mechanized. For example, suppose that the majority of the survey participants gave a positive answer to the following survey question in the classroom context: "Is it acceptable for ⟨*a professor*⟩ to share ⟨*a student's*⟩ ⟨*grade*⟩ with ⟨*the student's parent*⟩ ⟨*if the student gave her permission*⟩". The corresponding approved rule can then be expressed by the following

[2] http://yansh.github.io/papers/HCOMP/

[3] For norm disapproval score, we consider "No" answers.

[4] EPR is also known as the Bernays-Schönfinkel-Ramsey Class (Börger, Grädel, and Gurevich 2001).

conjunction:

$$
\begin{aligned}
&ctx = \mathsf{class} \wedge \\
&\mathsf{Professor}(sndr) \wedge \\
&\mathsf{ParentOf}(recp, subj) \wedge \\
&\mathsf{Student}(subj) \wedge \\
&attr = \mathsf{grade} \wedge \\
&\mathsf{Permission}(subj, attr, sndr, recp)
\end{aligned}
$$

If $R_1, \ldots, R_n$ are the logical rules obtained from the survey analysis, then the Allowed predicate can be defined by a universally quantified formula in EPR as follows:

$$
\begin{aligned}
&\forall ctx, sndr, recp, subj, attr. \\
&\quad \mathsf{Allowed}(ctx, sndr, recp, subj, attr) \Leftrightarrow \\
&\quad\quad (\ R_1(ctx, sndr, recp, subj, attr) \vee \cdots \vee \\
&\quad\quad\ \ R_n(ctx, sndr, recp, subj, attr)\ )
\end{aligned}
$$

That is, this formula states that Allowed captures exactly those CI flows that are admissible according to the accepted rules. Denote this formula by AllowedDef. We can then use this logical encoding to check automatically whether the derived privacy logic guarantees certain desirable properties. For example, suppose we wanted to check whether the derived rules guarantee that a student's grade cannot be shared without the student's permission. Then this property holds iff the following EPR formula is unsatisfiable:

$$
\begin{aligned}
&\mathsf{AllowedDef} \wedge \\
&\mathsf{Allowed}(\mathsf{class}, \mathsf{sndr}, \mathsf{recp}, \mathsf{subj}, \mathsf{grade}) \wedge \\
&\mathsf{Student}(\mathsf{subj}) \wedge \\
&\neg\mathsf{Permission}(\mathsf{subj}, \mathsf{grade}, \mathsf{sndr}, \mathsf{recp})
\end{aligned}
$$

The satisfiability of such formulas can be checked automatically using a decision procedure for EPR such as the one implemented in the theorem prover Z3 (De Moura and Bjørner 2008).

Evaluation

In our experiments we aim to:

- Evaluate how the metrics we propose can serve as indicators of the state of norms that have already been approved and whether users are satisfied with the socially derived privacy rule set;
- Test our automatic verification approach for consistency of the derived privacy logic.

Summary of crowdsourced data

Our survey design allows individuals to select norms according to their personal preferences and identify points of contention through formal verification techniques. As mentioned in the previous section, we presented users with a set of Yes/No/DMS questions. These questions were generated following the below template by traversing the values in CI parameters for sender, recipient, subject, attribute, TP space in the educational domain:

"*Is it acceptable for the ⟨sender⟩ to share the ⟨subject⟩'s ⟨attribute⟩ with ⟨recipient⟩⟨transmission principle⟩?*"

We summarize the user responses in Table 1. The "Total" column reflects the number of questions exceeding the respective NA thresholds for either 'Yes' or 'No'. There was overall consensus on approving or disapproving of 960 norms with NA greater than 50%. Not surprisingly, this number decreases as we increase the threshold to 66%. Next, we will discuss and provide examples of norms that fall outside the approved/disapproved/nondecided categories. Although these norms did not get a majority vote, they play a crucial role in our analysis of the overall sentiment of the crowd towards the norms that are accepted by the community. In other words, as in real life, privacy is rarely black or white, and often falls into a gray area. Thanks to our CI-based approach, we can examine this area in a systematic manner.

	Yes	**No**	**Total**
NA $> 50\%$	315	645	960 (68%)
NA $\geq 66\%$	115	300	415 (29%)
Yes = No			36 (2.6%)

Table 1: Summary of approved and disapproved norms across the surveys.

Approved and disapproved norms To provide the reader with an intuition for the type of norms that the community collectively agreed upon with the highest percentage of the vote, we list the CI parameter values for the approved and disapproved questions with NA $\geq 66\%$ in Table 2. Due to space restrictions, we only do so for the top five questions. However, we will release the complete dataset to allow the research community to explore our results in greater depth.

Note that in the Transmission Principle (TP) column of Table 2, we used numbers to represent transmission principles as follows:

1. *with the requirement of confidentiality*
2. *if subject is performing poorly*
3. *with a request from the subject*
4. *with subject's knowledge*
5. *with subject's consent*

Norms with no agreement For 36 questions, the respondents could not reach any agreement because the percentage of "Yes" and "No" answers was identical. We list the CI parameter values for the top ten of these questions in Table 3. These questions will require closer attention in subsequent surveys. At the moment, we can only speculate as to what caused the disagreement between the users. One reason could be that different users had different perceptions of the roles of individual actors such as professors, advisors, and the office of the registrar. In any case, the CI parameterization of the questions makes the exploration process more systematic as it can pinpoint the actual source of contention, which can be then addressed in subsequent surveys.

DMS Questions Only a small fraction of questions (8) have been classified as DMS by a majority of the users. We

Sender	Recipient	Subject	Atrribute	TP
Approved with NA $> 82\%$				
professor	graduate schools	student	attendance	5
registrar	parents	student	grades	5
professor	department chair	student	term papers	5
professor	registrar	student	participation	5
librarian	department chair	student	photo	5
registrar	graduate schools	student	name	5
Disapproved with NA $> 90\%$				
TA	classmates	student	grades	2
TA	classmates	student	transcript	1
TA	librarian	student	grades	3
advisor	classmates	student	transcript	1
registrar	classmates	student	transcript	3

Table 2: Summary of the CI parameters for top five approved and disapproved norms

Sender	Recipient	Subject	Atrribute	TP
registrar	librarian	student	grades	5
TA	graduate school	student	attendance	3
professor	advisor	student	term papers	2
professor	parents	student	contents online	1
registrar	librarian	student	photo	3
registrar	graduate schools	student	email	4
registrar	department chair	student	name	3
TA	IT staff	student	contents online	3
professor	advisor	student	term papers	1
classmates	parents	student	attendance	5

Table 3: Summary of CI parameters behind the questions that received equal percentage of Yes and No votes.

list those questions in Table 4. Upon closer inspection, one observes that these questions indeed may not capture meaningful interactions within the classroom context. For example, the question in the first row considers situations where *classmates* share the student's name with the student's *parents* if the student performs poorly. The other questions follow a similar pattern. This demonstrates that the crowd can help identify nonsensical CI parameter value combinations that our search space reduction techniques did not take into account.

Sender	Recipient	Subject	Atrribute	TP
classmates	parents	student	name	2
advisor	parents	student	name	1
advisor	parents	student	photo	2
advisor	parents	student	photo	3
TA	professor	student	name	4
IT staff	parents	student	photo	4
classmates	professor	student	grades	1
classmates	professor	student	grades	3

Table 4: CI parameter values of the questions classified as DMS

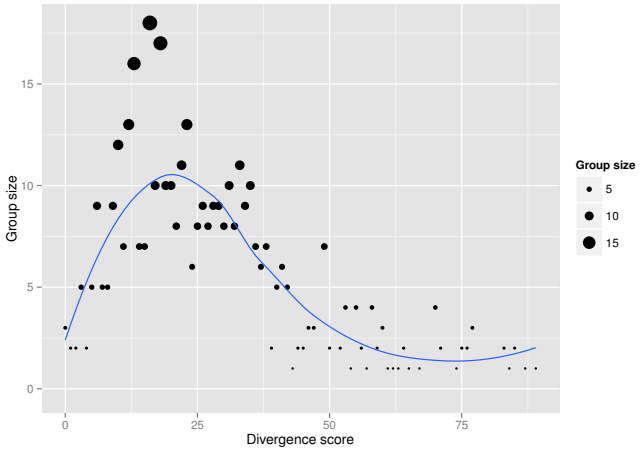

Figure 2: Scatter plot for divergence score

Quantitative Analysis

We now describe our quantitative analysis of the users' feedback, using the relevant metrics of norm approval (NA), divergence score (DS), and user approval score (UA).

User satisfaction While the NA threshold reflects a lower bound on the number of users that approve or disapprove of a norm, the DS score serves as an indicator how the users feel about the approved norms. A lower DS score indicates a higher agreement within the surveyed community about the approved norms.

We analyzed the different NA thresholds and how these threshold choices affect the users' approval and divergence scores. We focused first on the two thresholds of 50% and 66%. Figure 3 depicts two boxplots for all users across all questions for the two NA thresholds. Both populations look very similar, with a DS mean of approximately 25, which suggests that, at both thresholds, individual views on approved norms aligned with those of the overall community.

To verify the difference in means of these two populations, we ran a one-way ANOVA to test a *null* hypothesis that there is no difference between the populations of means under different thresholds. We can reject the null hypothesis with significance level $p = 0.000165$ ($p < 0.05$). A Tukey HSD test identified that using the 66% threshold increases the DS score by 4.8%.

In other words, this shows that the 66% threshold results in a higher disapproval among users with regards to their expressed privacy expectations.

To further visualize this result, the scatter plot in Figure 2 depicts the number of users with the same DS score across all questions for an NA threshold of 66%. The plot suggests a large concentration of respondents with a relatively small DS. This means that, overall, the users in our polls are satisfied with the privacy rule set that is determined by the specific NA threshold.

Users' tendency to approve or disapprove of norms. We calculated a combined DS for all possible NA thresholds (0% to 100%) and normalized it by the number of total

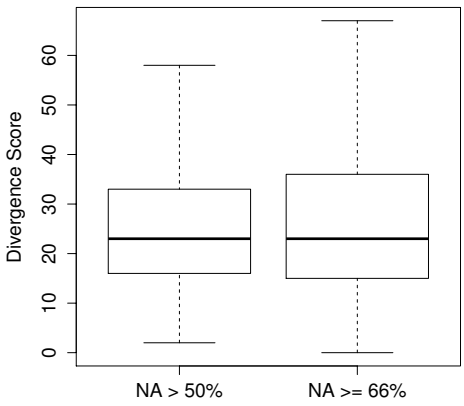

Figure 3: DS for 50% and 66% NA thresholds: depicts how individual users' DS varies with the respective NA threshold values.

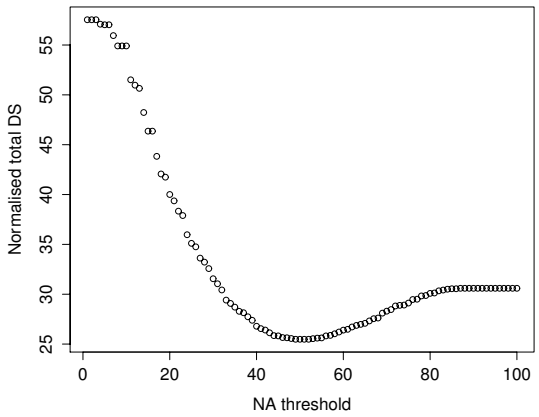

Figure 4: Total DS across all possible thresholds: for each NA threshold we calculated and aggregated the DS score for each of the users. The final number was normalized by the number of the users.

users that had taken the survey. This normalization provides us with the combined DS score of all users per threshold. The resulting data is provided in Figure 4 and shows that, when the threshold is at its minimum, the DS is at its maximum. Recall that the DS represents the level of dissatisfaction among users. We can therefore interpret this result as follows: when the threshold is low, more questions are approved, meaning that a significant number of privacy rules that users prefer to disapprove are actually accepted by the overall community. The lowest DS values are in the 40% to 60% NA threshold range. The best candidates for an actual threshold choice, for this specific population based on their feedback, therefore seem to lie in that range. Interestingly, the DS converges around the 35 mark from 66% to 100%. This shows that, in our polls, more people opt to disapprove norms rather than approve them.

Individual privacy expectations vs social norms Figure 5 shows that there is a linear relationship between UA and DS for individual users for a 66% NA threshold. Linear regression analysis confirms this (r^2 = 0.87, formula: $DS = 0.69 * UA + 2.044$). The 66% NA threshold makes it hard to approve privacy rules; users with a very high UA score will often be disappointed, thus having higher DS. Conversely, users that have a lower UA are more likely to agree with the community rules. We can observe a similar pattern with an NA threshold of 50% on Figure 6; however, relative to the 66% threshold, user satisfaction is slightly higher as more privacy rules are approved on average.

Verification experiments We also evaluated the effectiveness of formal verification technology to analyze the consistency of the derived privacy rules. We used the theorem prover Z3 to check whether the crowdsourced rules guarantee certain privacy properties by encoding both the rules and the properties into EPR as described earlier. Specifically, our goal was to assess whether we can use the theorem prover Z3 to automatically check the consistency between the rules that we derived from the crowd-sourced data for a chosen threshold, on the one hand, and to check for consistency violations, on the other hand.

We focused our attention on two specific consistency properties:

1. Semantic consistency of rules. This property specifies that the information flow of each disapproved norm is indeed excluded from the flows that are allowed according to all approved norms. Note that this property is not trivially satisfied as the approved and disapproved norms are not necessarily mutually exclusive. In particular, the roles of a context are not guaranteed to be disjoint, e.g., an actor in the classroom context may be both a department chair and a professor. Thus, we may have situations where a specific flow is approved if the sender is a professor but disapproved if the sender is a department chair. Such inconsistencies hint at hidden assumptions of the survey participants that are not adequately reflected by the formal privacy rules. Our verification approach allows us to detect such inconsistencies and subsequently eliminate them by refining the formal rule model and the survey questions appropriately.

2. Consistency of transitive flows. This property specifies that the approved norms are transitively closed. A violation of the transitivity property hints at a possible mismatch between the survey participants' privacy expectations and the logical implications of their individual choices regarding which privacy norms should be approved. Using our verification approach, we are able to detect such violations (respectively, prove their absence) for arbitrarily long sequences of information flows.

In the following, we describe the experiments we conducted to check for each of these two properties as applied to the set of norms that we derived from our survey data.

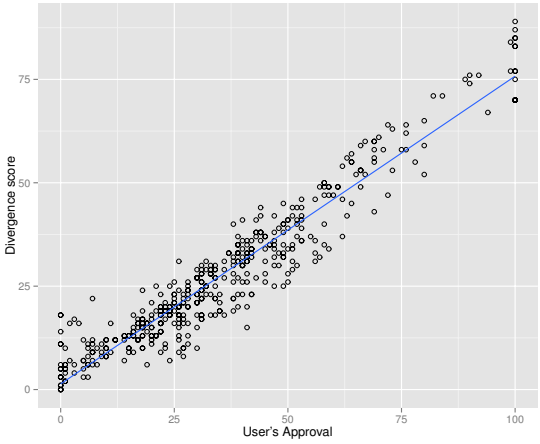

Figure 5: Scatter plot of Users' Approval and Divergence Score for each user for NA threshold 66%.

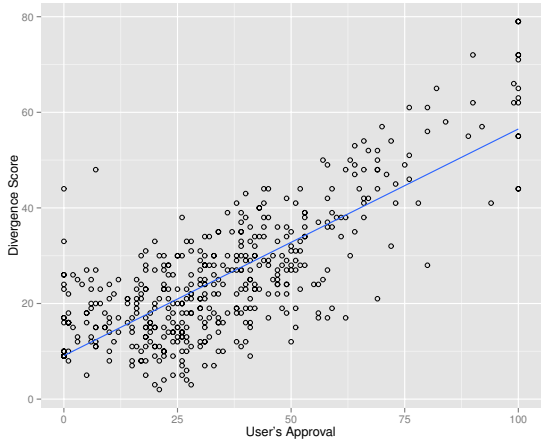

Figure 6: Scatter plot of Users' Approval and Divergence Score for each user for NA threshold 50%.

Note that in both experiments the entire verification process, including the norm extraction, the logical encoding of the norms and properties, and their verification, was fully automated.

All the experiments were run on a laptop computer equipped with an Intel Core I5 CPU at 2.67GHz and 4GB RAM running Ubuntu Linux. The running time for each of our experiments was less than 5 seconds. The memory consumption was negligible.

Detecting semantic inconsistencies of norms For our experiment, to detect semantic norm inconsistencies we chose the 50% threshold to determine which norms are approved according to the crowd-sourced survey data. For this threshold, as depicted by Table 1, 315 of our total 1411 norms were approved. We then encoded these approved norms into an EPR formula and used Z3 to check for each of the 1096 remaining disapproved norms whether the corresponding information flow was indeed prevented by the approved rules. Each disapproved norm was checked by sending a separate satisfiability query to Z3.

Intuitively, semantic norm inconsistencies can only arise if an agent takes on more than one role in a context at the same time. We confirmed this intuition by conducting an experiment where we verified the absence of inconsistencies under the assumption that all roles are pairwise disjoint. Indeed, under this assumption we were able to prove that 100% of the disapproved norms were consistent with the rules for the approved norms.

To detect actual semantic norm inconsistencies, we considered a model that took the relationships between the different roles in a classroom context into account. For example, a TA may also be a student and a department chair is always a professor. With the realistic model, we detected that 138 of the 1096 disapproved norms were not ensured by the approved norms. For example, one of the violated disapproved norms pertained to a professor sharing a student's test result with other students. Such an information flow was permitted by one of the approved norms, which allowed a professor to share a test result with a TA. Since a TA may also be a student, the disapproved norm was indeed violated.

There are a number of possible ways in which such violations could be resolved (e.g., by refining the privacy rules or domain ontology). These are outside the scope of this paper. The focus of our experiment was to demonstrate that we can automatically detect all such violations, or alternatively prove their absence.

Detecting inconsistencies in transitive flows The final experiment was designed to check for inconsistencies due to transitive flows. Similar to the previous experiment we encoded the logic into an EPR formula and used Z3 to check for any violations of the transitivity property. The transitivity property involves reasoning about arbitrarily long chains of information flows. This means that for a specific set of approved norms, the number of concrete chains of information flows that are consistent with the rules but violate transitivity may be infinite. However, we observed that for any specific violation, there always exists a *similar* violation involving a chain of bounded length. This means that all transitivity violations can be classified by a *finite* set of small violations. This observation allowed us to exhaustively enumerate all types of transitivity violations for a given set of approved rules. To do so, we used Z3's model generation capability to generate models that witness a small violation of transitivity.

For the 66% threshold, where 115 of our total 1411 norms were approved, we automatically detected 59 transitivity violations. On closer inspection, we found that one such violation was the result of the following two approved norms:

1. *A TA is allowed to send information about a student's attendance to a professor if the student is performing poorly.*
2. *A professor is allowed to send information about a student's attendance to the department chair if the student is performing poorly.*

However, a TA was not allowed to send the attendance in-

formation directly to the department chair, leading to a violation of transitivity. The approval rate of this rejected norm was only 17%. Contrasted with the high approval rates of more than 66% for the two approved norms involved in the above transitive flow, this discrepancy hints at a possible violation of the actual privacy expectations of the users.

Discussion

The theory of contextual integrity serves as a tool for understanding and reasoning about the appropriateness of context-specific information flows. Combined with the methodology of crowdsourcing, it offers a gateway into users' collective privacy expectations, as we are able to ask large numbers of people about their sharing preferences with regards to particular information flows, on the one hand, and translate their preferences into contextually appropriate and hence collectively acceptable privacy norms, on the other hand. Our evaluation suggests moreover that we can analyze these parameters and detect those privacy norms that users are more likely to approve and care about.

Broader applicability

For demonstration purposes, we limited the application of our framework to the educational context. In principle, this same approach could easily be applied to other contexts or a much broader set of actors, attributes and transmission principles. In addition, the same methodology can be applied in an incremental fashion when new actors, attributes or transmission principles are added to the context definition.

Limitations and extensions

We believe our results provide a solid indication that the method we chose, namely discovering informational norms on the basis of crowdsourcing, has merit. Nevertheless, we would like to acknowledge a few limitations of our analysis that could be improved upon in future work.

First, we note that, due to the size of the question space, the user is presented with a relatively large number of questions (88), the answering of which requires a substantial cognitive effort which may result in survey fatigue. In principle, this number can be further reduced by categorizing and prioritizing the questions based on their importance to the users. For example, users might care more about privacy policies that relate to the flow of their personal, rather than more generic data. For less important attributes, we can introduce default values.

Secondly, the number of answers per question is relatively small (32) which allows small variations to have a large "ripple-effect" in the selected norms. In the current setting, every vote counts. To counteract this limitation we introduce the 66 percent threshold to ensure a meaningful majority when selecting the norm. In future work, we plan to increase the number of participants, though, realistically, due to budget constrains, the number will always be low compared to the number of users in existing online social networks. We hope that, by making our work public, we will be able to attract industry collaborations to refine our methodology in a real-world setting.

Finally, although prior work (Lin, Amini, and others 2012; Martin 2014; Ismail, Ahmed, and others 2015; Kandasamy, Curtis, and others 2012) offers early validation of crowdsourcing methodologies, suggesting that large-scale surveys can indeed be effective for the discovery of social norms, we realize that the results of individual users might differ in reality. This is something that we would like to address in future work by integrating our methodology into a real-world system.

Despite these limitations, we believe that the use of crowdsourcing for the purpose of the discovery of social norms is useful and may ultimately serve as the basis for a consistent and reliable privacy logic that is supported by the majority of a system's users.

Related Work

In this section we acknowledge important prior and adjacent work that is related to ours.

(Sadeh et al. 2013) proposed a framework as part of the Usable Privacy Project[5] that capitalizes upon natural language processing (NLP), privacy preference modeling, and crowdsourcing to capture privacy policies used by websites and translate them into simplified models. The models are used to help users in making privacy-related decisions through the analysis of historical policy trends, detecting inconsistencies between policies and identifying key features relevant to the user's privacy profile. While the main goal of the project, namely "[empowering] users to more meaningfully control their privacy," resonates well with our efforts, we are primarily interested in providing tools to policymakers for discovering both contextually appropriate as well as contentious information flows. Furthermore, our approaches differ significantly when it comes to the use of crowdsourcing, formal modeling and verification methods.

One effort that is quite close in spirit to our own is (Lin, Liu, and others 2014) which seeks to ease the burden on users when tailoring privacy policies on mobile apps to accurately reflect their privacy preferences. The system clusters users according to their willingness to share information with app providers and configures settings on future apps based on the position in a cluster. Relevant differences are (i) that it applies to a dyadic relationship between the user and app provider, and (ii) that it seeks to model preferences while our work aims to translate social norms.

Similarly, (Toch 2014) has proposed the SuperEgo system, which uses crowdsourcing to enhance location privacy management in mobile applications. SuperEgo uses the perception of the crowd to predict the privacy preferences of an individual. The system relies on a crowd-opinion model and a mixture of decision-making strategies to classify the information as private or not. Although this work is conceptually similar in that it uses crowdsourcing to infer relevant privacy policies for the user, it is limited to a location-based privacy context. As noted by the author, the CI framework is more expressive and capable of capturing privacy-rules in a range of different contexts.

[5] https://usableprivacy.org

In (Lin, Amini, and others 2012), AMT was used for crowdsourcing users' privacy expectations with regards to the access of applications to resources available on a phone. Although this work, as stated by the authors, only considers a "narrow construct" in the privacy expectations domain, namely that of "people's mental models of what they think an app does and does not do," it serves as a great motivation for our approach as it shows that users' privacy expectation tend to vary based on context and often differ from what is formally allowed by a privacy policy.

Summary and Conclusion

In this paper we described a framework for crowdsourcing privacy norms based on the theory of contextual integrity. Our evaluation demonstrates that the combination of crowdsourcing and CI's privacy model allows us to distinguish between socially acceptable and contentious informational norms, on the one hand, and to identify those norms that make little sense, or require further examination, on the other hand.

To facilitate this effort, we introduced several indicators that can be used to produce an estimate of the overall community's privacy expectations towards a set of information flows within a given context. More specifically, NA allows us to measure the approval score for each norm; UA to reflect the number of norms approved by individual users; and the DS metric allows us to measure the overall satisfaction of users with regards to the collectively agreed upon privacy norms. These indicators can be used to fine-tune privacy policies on an ongoing basis to ensure that the majority of users actually support the enforced logic. Furthermore, we leverage formal verification techniques to detect inconsistencies in approved sets of norms that can then be eliminated by refining the survey questions.

To evaluate our metrics, we conducted an extensive survey on AMT with more than 450 participants and 1400 questions. Our methodology allowed us to discover norms that have majoritarian consensus among users, discard norms that do not make any sense in a given context as well as detect norms that require further examination. Our results leave us optimistic about the feasibility of a full-fledged information system that operates based on the design principles of crowdsourcing, formal verification, and contextual integrity. Future work includes an in-depth investigation into more elaborate approval and divergence metrics, an extension of our design to handle inter-domain privacy rules as well as the release of a prototype system based on privacy norms discovered using the methods we have proposed.

Looking ahead, our work paves the way towards developing information systems that operate on the foundation of substantive privacy rules that reflect the rough consensus of given communities. These could include communities across the domains of education, health, or more general social domains. The mechanisms we have developed for extracting, expressing, and validating a set of common rules could be integrated into such systems. Discovering a set of common rules through crowdsourcing, e.g. as we have done through AMT, can be viewed as an efficient bootstrapping first move to populate a system at its budding stage. Incorporating these mechanisms into social information systems allows us to elicit user feedback on an ongoing basis, which then ultimately enables our system to respond to evolving community norms and standards.

References

Barth, A.; Datta, A.; et al. 2006. Privacy and contextual integrity: Framework and applications. In *Proc. of the IEEE S&P*, 15–pp.

Börger, E.; Grädel, E.; and Gurevich, Y. 2001. *The classical decision problem*. Springer Science & Business Media.

Chowdhury, O.; Gampe, A.; et al. 2013. Privacy promises that can be kept: A policy analysis method with application to the hipaa privacy rule. In *Proc. of the ACM SACMAT*, 3–14.

Computing Community Consortium. 2015. Privacy by Design. http://cra.org/ccc/visioning/visioning-activities/2015-activities/privacy-by-design/.

Criado, N., and Such, J. M. 2015. Implicit contextual integrity in online social networks. *arXiv preprint arXiv:1502.02493*.

De Moura, L., and Bjørner, N. 2008. Z3: An efficient SMT solver. In *Proc. of the TACAS*. Springer. 337–340.

Ismail, Q.; Ahmed, T.; et al. 2015. Crowdsourced exploration of security configurations. In *Proc. of the 33rd ACM CHI*, 467–476.

Kandasamy, D. M.; Curtis, K.; et al. 2012. Diversity within the crowd. In *Proc. of the ACM CSCW*, 115–118.

Krupa, Y., and Vercouter, L. 2012. Handling privacy as contextual integrity in decentralized virtual communities: The privacias framework. *Web Intelligence and Agent Systems* 10(1):105–116.

Lin, J.; Amini, S.; et al. 2012. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proc. of the ACM UbiComp*, 501–510.

Lin, J.; Liu, B.; et al. 2014. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Proc. of the SOUPS*, 199–212.

Martin, K. 2014. Privacy notices as tabula rasa: An empirical investigation into how complying with a privacy notice is related to meeting privacy expectations online. *Journal of Public Policy & Marketing* 34(2):210–227.

Nissenbaum, H. 2010. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books.

Posner, R. A. 1977. The right of privacy. *Ga. L. Rev.* 12:393.

Sadeh, N.; Acquisti, A.; Breaux, T. D.; et al. 2013. The usable privacy policy project. Technical report, CMU-ISR-13-119, Carnegie Mellon University.

Toch, E. 2014. Crowdsourcing privacy preferences in context-aware applications. *Personal and Ubiquitous Computing* 18(1):129–141.

Westin Alan, F. 1967. *Privacy and Freedom*. Atheneum. New York.