Tackling Compliance Deficits of Data-Protection Law with User Collaboration – a Feasibility Demonstration with Human Participants

Thorben Burghardt, Erik Buchmann, Klemens Böhm
Institute for Program Structures and Data Organization,
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
firstname.lastname@kit.edu

Jürgen Kühling, Simon Bohnen, Anastasios Sivridis
Chair for Public Law and Real Estate Law,
University of Regensburg
Regensburg, Germany
firstname.lastname@jura.uni-regensburg.de

***Abstract*—In the recent past, there have been frequent reports on privacy violations by service providers on the Web. The providers are overstrained with the legal implications of processing personal data. Data-protection authorities in turn are overburdened with the enforcement of the regulations. Users themselves typically cannot identify those violations, due to missing expertise in data-protection law. In this paper we propose and evaluate *CAPE* (*C*ollaborative *A*ccess to *P*rivacy *E*nhancement), an approach that makes data-protection law accessible to all parties involved in the processing of personal information. To this end, we transform legal expertise on data protection into intuitive questions that anyone can answer. *CAPE* is 'Web 2.0', in the sense that individuals answer the questions they can, and they benefit from the answers of others. To identify violations, we compare the answers to answer patterns defined apriori that indicate a violation. The main innovation is the combination of Web 2.0 functionality with the structured approach (sequences of closed questions in particular) lawyers use to identify violations. In extensive user studies, we show that users can identify 81% of those violations legal experts find. Further, individuals answer our questions with a high degree of agreement, independent from their background knowledge.**

***Keywords*-data protection; compliance deficit; collaboration; user study; e-government**

I. Introduction

In the recent past, the press has reported on dozens of data-protection violations[1]. Many of them would have been avoidable – if existing data-protection regulations had been enforced. [1] has shown the drastic extent of this compliance deficit and its implication for the individual.

We see two main reasons for the deficit: First, the resources of the authorities monitoring abidance by data-protection law are limited, while the number of data-collection activities on the web is daunting. To give an intuition, 6 assistants of the data-protection commissioner have to supervise the Internet presence of 100,000 companies in one German state. Second, it requires expert knowledge to understand today's data-protection acts. For example, unspecific terms like 'appropriateness of usage' need to be interpreted for the context under consideration. Further, there is a huge amount of different privacy norms ($> 1,000$ in German legislation alone). It is challenging to overcome the compliance deficit of data-protection acts.

In this paper we introduce *CAPE* (*C*ollaborative *A*ccess to *P*rivacy *E*nhancement), an approach to tackle the compliance deficit of data-protection law. Our approach targets at interested individuals, data-protection authorities, enterprises, certification authorities, and consumer-protection agencies. It lets individuals evaluate the privacy practices of online services (web shops, discussion forums, search engines, etc.) in an intuitive way. To this end, legal experts have come up with a taxonomy of intuitive, privacy-relevant questions, and with answer patterns on these questions indicating a data-protection violation. Persons without legal knowledge can answer the questions easily. The taxonomy describes relationships between questions. Two questions are related if (i) one requires more detail knowledge than the other one, or (ii) one is asked only if the other one has been answered in a specific way. Finally, *CAPE* matches the patterns against the answers provided by the users in order to identify privacy violations. The collaboration aspect of our Web 2.0 approach allows individuals to answer the questions they can, and they benefit from answers of others. Further, anybody using the Web can benefit from relatively few contributors. We have taken into account that some violations depend on complex patterns, and that some violations can be identified in several ways. For example, users can observe the registration process, browse the privacy policy, test for cookies and web-analytics services etc.

Example 1: Think of the legal aspects of *data acquisition*, *data forwarding* and *consent*. Questions related to these aspects include q_1 ='Does the company ask for personal data?', q_2 ='Does the company forward data to non-EU countries?', q_3 ='Will the data be sent to Argentina, Guernsey, Isle of Man, Canada, or Switzerland?'[2] and q_4 ='Did you have to consent to the privacy policy?'. A user community answers these questions for a number of service

[1] e.g., http://www.idtheftcenter.org, Breach Database

[2] The EU offers a list of countries having a level of data protection comparable to the one of EU member states. Further, in this example we leave aside US enterprises that conform to the Safe Harbor Agreement.

1530-1354/10 $26.00 © 2010 IEEE
DOI 10.1109/CEC.2010.34

providers. The pattern *<yes, yes, no, no>* models a violation against data-protection law in all EU countries. If a significant number of answers matches this pattern for a certain provider, *CAPE* flags a violation.

Note that this approach does not prove that a provider complies with the law. However, this is the first proposal that promises to deal with the compliance deficit of existing privacy regulations. It can be extended to any new violation, law and context considered. In particular, our approach has the potential to identify a large number of privacy violations with the help of interested users, it allows enterprises to check internal processes against the violations modeled, and it lets the data-protection authorities enforce law much more efficiently. Thus, its social impact is high.

Once having designed the approach, implementing it is relatively straight forward. Thus, our description is conceptual and not at a technical level.

In this article, we make the following contributions:

- We motivate and describe *CAPE*, our Web 2.0 approach to enable a community of individuals to detect privacy violations collaboratively.
- We propose a methodology that guides legal experts through the process of mapping legal norms to a set of intuitive questions. Given this, we build a taxonomy of questions for the German data-protection law for online services. As the EU harmonizes data-protection law between the member states, e.g., in EU Directive 95/46/EC, the taxonomy is applicable in all other EU states with slight modifications.
- To find out if a community of users without legal expertise can identify data-protection violations using *CAPE*, we have carried out a user study with 77 participants from different social groups. We show that these user communities can find 81% of a wide range of violations legal experts find.

Paper structure: Section II reviews related work. Section III introduces our approach. Section IV features the methodology of our study and the results, Section V concludes.

II. Related Work

In this section we review related work. These are collaborative privacy approaches, privacy enhancing technologies (PETs), tools supporting the creation of privacy policies, and early approaches applying natural language processing.

Collaborative Privacy Approaches: The e-commerce community has intensively studied privacy aspects of collaborative systems [2], [3]. However, there are only few approaches using the wisdom of a community to deal with data-protection issues. [4] proposes to assign privacy levels to individual privacy preferences based on the community consensus. [5] describes a study where a community of users creates a Web 2.0-folksonomy of potential data-protection problems. Such problems may arise from RFID-tagged items, geo-locations or, relevant for our context, websites. The authors propose to use the folksonomy to notify users of privacy-threatening sites or objects. [6] uses annotations to cluster users according to their privacy attitude. Preferences from users with a similar attitude, i.e., users from the same cluster, allow to predict privacy preferences for providers the user has not yet stated a preference for.

Our objective is different from the ones of the approaches described: Our basis is the implementation of legal expertise and the systematic identification of privacy violations. This means that law defines whether a violation exists or not. This is in contrast to subjective feelings of individuals on what should be private. The advantage of our approach is that a user, having correctly identified a violation, has a legal basis to claim his rights, or to bring the violation to court.

PETs: There exists a large variety of PETs for many application domains. A prominent approach is P3P [7]. P3P allows to define a privacy policy in a machine readable way, and clients can use tools like PrivacyBird[3] to test if the policy of a provider matches their individual privacy preferences. However, only 3.5% of the service providers on the Internet use P3P [8]. Further, the expressiveness of P3P is limited, and P3P cannot provide all information required by EU law. For instance, it is impossible to express all circumstances relevant for cross border data forwarding. Our community-based concept promises to fill this gap.

Creating Privacy Policies: There exist tools to create privacy policies that match legal requirements, e.g., the OECD Privacy Statement Generator[4] or the Privacy Policy Generator[5]. If most institutions use such tools, this might increase the quality of privacy policies. However, the core problem is that the policies generated state what the policy creator wants to express – not what he actually does. For example, saying 'no automated processing of personal data' and using Cookies are in conflict. We in turn validate policies that already exist and check for possible conflicts. Further, our concept is not limited to privacy policies but includes, say, the registration process, practices on giving consent or on using web-statistic tools, etc.

NLP techniques for policy parsing: [9] adapts natural language processing to privacy policies, as follows. They try to extract the privacy practices of a provider from his policy automatically. This requires an intensive preprocessing of each privacy policy. In more concrete terms, humans have to transform the policies in a format which their rule engine understands. Given the huge number of Internet providers (and the current state of the art of NLP), we deem this impracticable.

[3] AT&T, PrivacyBird, www.privacybird.org

[4] http://www.oecd.org/document/39/0,3343,en_2649_34255_28863271_1_1_1_1,00.html

[5] http://policygenerator.net

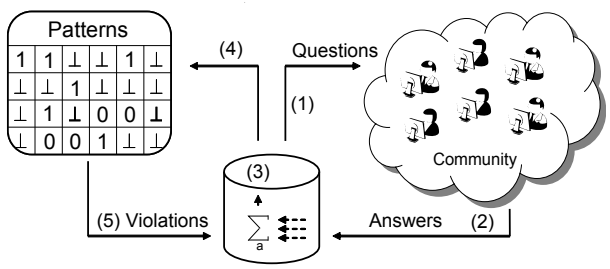

Figure 1. The *CAPE* Approach

III. The *CAPE* Approach

In this section we describe our approach. Figure 1 gives an overview. We assume that experts have come up with a taxonomy of questions and have defined patterns of answer combinations which represent data-protection violations. Questions and patterns are stored in a database. Then the system asks the users questions from the taxonomy (1). *CAPE* collects the answers (2). In the following, we will refer to answering the questions specific for one provider as a *review*. Using the answers, *CAPE* builds the community consensus (3) and compares it to the violation patterns (4). Last, it stores the violations identified (5).

We have different target groups in mind, in particular Internet users, data-protection authorities, certification institutions, and individuals responsible for data protection within a company.

In what follows, we first specify our requirements (Section III-A). Second, we introduce our system architecture (Section III-B). We will describe its components on a logical level; its implementation makes use of off-the-shelf technology and has not been difficult. Third, we present a methodological framework to map legal expertise to a taxonomy of questions and answers (Section III-C).

A. Requirements

Besides technical requirements, e.g., extensibility, performance, etc., our system has to fulfill functional requirements. These requirements stem from common demands on privacy approaches, from the target groups we want to address, and from the fact that the legal background is continuously adapted to new privacy issues.

Transparency (R1) Users must be able to see which answer has led to the detection of a violation. Furthermore, the legal background for the violations detected must become clear.

Support for different user groups (R2) We have to consider that our target groups have different perspectives on privacy issues. For instance, users can have different insight into the company, can be registered users or not, etc. Further, they have different motivations to identify violations, e.g., making experiences public or testing internal business processes.

Model Independence (R3) The legislator continuously adapts regulations to cover new privacy issues. Thus, our models of the legal expertise, of the reviews by the users, and of the identification of violations must be independent from each other.

Stochastic Guarantees (R4) Induced by the different interests of the user groups (R2) and the continuously changing legal framework (R3), (R4) requires to deal with mistakes and with insufficient or contradicting reviews. We need a stochastic model that ensures confidence bounds for the violations identified.

B. System Model

This section describes our system model. We introduce the notion of *questions*, *answers* and *patterns* first.

Questions Norms contain unspecific statements like 'Data collection [...] shall be admissible in line with the purposes of a contract [...].' that need to be interpreted. Legal experts map such norms to concrete questions for specific use cases, e.g., 'Does a web shop ask for more data than needed for billing and shipping?'.

Answers The users contribute by answering the questions. The answers represent the experiences and the knowledge of a community of users.

Patterns We let legal experts model privacy violations as patterns of possible answers. See Example 1 for a pattern leading to a violation.

In order to realize the process model of Figure 1, we propose a system architecture consisting of a taxonomy component, a collaboration, and a detection component.

Taxonomy Component: The taxonomy reflects that questions can require different levels of detail knowledge, e.g., knowing if a provider is using cookies vs. knowing the kind of cookies. Furthermore, the taxonomy considers (together with the detection component) that the same violation can be modeled and identified in different ways, depending on the perspective of the users. To this end, experts can define hierarchical relationships between the questions. Thus, relationships can have two types: *Dependency* relationships identify questions that are asked only if other questions have been answered in a specific way. *Abstraction* relationships model questions on different levels of knowledge.

Example 2: Think of the questions q_1 = 'Does the provider use cookies?' and q_2 = 'Does the provider declare the storage period of the cookie?'. Regarding q_1, a user can get this information from her browser. A dependency relationship specifies that q_2 can be omitted if q_1 = 'no'. Now consider q_3 = 'Does the provider use session cookies?'. q_3 is more specific than q_1. An abstraction relationship specifies that q_1 is 'yes' if a user answers q_3 = 'yes'.

We store both questions and relationships in a relational database.

Collaboration Component: This component administers the answers from the users. Users can have different perspectives (roles), e.g., 'data-protection commissioner',

'individual responsible within a company' or 'Internet user'. 'Legal experts' have a special role, since they are the only group that can define questions and violations. The collaboration component relates each answer to the user who gave the answer, the query answered, and the service provider. We store the user information and the answers they give in the database.

Answers can vary, e.g., if some persons use ad or cookie blockers and others do not. The collaboration component decides which questions have to be answered by additional users, and it allows to detect users with outlier answers. Further, it allows to define a function to derive the community opinion, i.e., one single answer representing all users.

Detection Component: This component decides if a provider commits one or more violation. The detection component has to consider four aspects: First, the set of incomplete, misleading or contradicting answers can be insufficient to come to a conclusion that is statistically significant. For instance, we calculate the inter-rater reliability, questions difficult to answer, the probability of a randomly given answer, etc. Second, questions can have a different degree of abstraction. Third, the same violation can be detected from different perspectives (e.g., internally/externally). Finally, some laws require k out of n criteria.

We have decided to implement a simple detection mechanism: We let experts model violations as patterns. A pattern is a vector of length $\#questions$ of the taxonomy, which we store in our database. It consists of <value, logical operator>-pairs. The 'value' can be true, false or a number. The logical operator can either be "don't care" ($\perp$ in Figure 1), i.e., the answer will be ignored, or '=' for binary questions respectively $=, >, <, \leq, \geq$ for numeric values. The operator compares the 'value' to the answer of a question. If, taking the review of a provider, each element of a vector evaluates to 'true', we have identified a violation.

Example 3: Think of the question q_1 = 'Does the provider use cookies?' and q_2 = 'Does the provider state to use cookies in his privacy policy?'. A pattern that tests for a violation would consist of <q_1, true, '='>,<q_2, false, '='> , leaving the values for all other questions "don't care".

Note that we can apply patterns to the answer of a single user or to the community opinion, i.e., an aggregate of the answers of all users. All elements of a vector are implicitly connected with a logical AND. In order to model that the same violation can be identified in different ways, the legal experts can come up with multiple vectors. The vectors are connected with a logical OR. Once a violation v is identified, we store it in the database.

We have consciously decided to keep our approach simple, in order to enjoy several advantages. The approach fulfills our requirements (Section III-A): It is evident which particular combination of answers has identified a violation (R1). The collaboration component distinguishes answers from different user groups (R2). The separation of the components (R3) allows for an easy adaption to new regulations. The relational storage we use gives way to easy deployment of standard statistical tools (R4).

C. Methodological Framework

Our methodological framework is inspired from common legal methodology (syllogism, cf. [10]), which consists of consecutive questions, e.g., "Who wants to have what from whom and for which reason?" Together with legal experts we have defined an iterative process. It guides others, including experts who want to extend *CAPE*, to transform the legal background of known data-protection violations into questions. The questions can be answered by non-experts and allow to identify violations. The framework has 6 steps:

1. Legal Basis: The first step specifies the legal basis for the kind of provider considered, i.e., it identifies the data-protection acts which the providers are subject to. For example, in Germany most services available on the Internet are regulated by the Federal Data-Protection Act (Bundesdatenschutzgesetz, BDSG) and the Telemedia Act (Telemediengesetz, TMG), which implement the EU Directive 95/46/EC.

Example 4: To answer if the TMG can be applied, we have to know from the users, among others: q_1 = 'Does the provider offer Internet access only?', and q_2 = 'Is the service a web forum that is not moderated?' If q_1 is answered 'yes', the law for Internet access providers is the relevant one, not the TMG. If a forum is carefully moderated (q_2), this indicates a professional journalistic background of the provider, according to German law, i.e., another law is relevant. Otherwise, the TMG can be applied.

2. Perspective: User can detect violations from an internal or from an external perspective (cf. Section III-A). This step identifies the perspectives to be considered.

Example 5: Think of q_1 = 'Does the provider declare to create pseudonymized profiles in the privacy policy?' and q_2='Does the company correlate pseudonymized data with data sources such that this correlation gives way to a personal identification?'. While both questions can be answered from an internal perspective, e.g., from a person responsible within a company, the second question cannot be answered from an external perspective, e.g., an Internet user.

3. Concretion: This step deals with unspecific formulations like 'reasonable', 'appropriate' or 'should', according to the legal basis (Step 1), i.e., the kind of service provided. Two aspects have to be considered: (i) given an unspecific formulations, does a violation exist for that kind of service (latitude of judgement), and (ii) if a violation exists, what are the legal consequences (judgement evaluation)?

Example 6: Section 13 Paragraph 6 TMG says that as long as anonymous usage is reasonable, the provider has to offer this feature. The meaning of 'reasonable' is a latitude of judgment. In the case of web shops, it is common practice to browse catalogues

anonymously. Thus, it is reasonable to offer such an option, otherwise the provider might violate the law.

4. Real-World Instances: Law is generic, i.e., it uses abstract terms to subsume real-world concepts. This step maps legal terms to concrete real-world instances.

Example 7: The TMG requires a provider to declare the use of automated processes, but it does not specify any concrete technology. For instance, the questions q_1 = 'Does the provider use Cookies?' and q_2 = 'Does the provider use web statistic tools?' map the term 'automated processing' to real-world instances on the Internet.

5. Interpretation: This step again deals with the generic characteristics of law. Norms do not explicitly state for each circumstance and technology which behavior is correct. To identify the legal consequences for a given context, this step interprets the wording, history, and intention (telos) of each norm applied, and the relevant circumstances.

Example 8: The TMG requires a provider to declare the storage time of cookies, but it does not say anything regarding false declarations explicitly. Nevertheless, as the intention of the legislator is to ensure transparency, a false storage time is a violation.

6. Implementation: This step consolidates the questions developed in the earlier steps. The questions must be as simple as possible, e.g., simple yes/no-questions. This step removes duplicate questions, decomposes questions that are too complex and adapts the violation patterns accordingly.

As said, the process has to be applied iteratively until all questions are so simple that anyone can answer them.

IV. User Study

In this section we describe the setup, the procedure and the results of a user study that evaluates our approach. The legal background relevant is the German Federal Data-Protection Act (BDSG) and the Telemedia Act (TMG). Using our methodological framework, we have created a taxonomy of questions for service providers on the Internet from the private sector. The participants have the perspective of Internet users, i.e., they browse the privacy policy, register at the provider, analyze cookie usage, etc. However, they have no insider knowledge of the providers. We measure if (Q1) *CAPE* allows participants to answer the questions and review the providers in reasonable time, (Q2) the answers converge to a consensus and (Q3) the consensus allows to identify violations correctly. Finally, we evaluate how different user groups with a different background knowledge on data protection perform (Q4).

Our complementary website[6] contains excerpts of the taxonomy, a manual of *CAPE* as well as screenshots.

A. Study Setup

Our study setup is based on six key design decisions:

[6]http://privacy.ipd.kit.edu

Participants: We have decided for three groups of participants with different levels of education, experiences and interests: (i) pupils from a German Gymnasium (pup), (ii) undergraduate students from technical disciplines (cs), and (iii) law students that are familiar with legislation (law). Altogether 77 individuals between 13 and 29 years (avg. 22) have taken part in our study. 49 participants have been male, 28 female.

Incentive: There should be an incentive stimulating participation. It must not induce the user to answer a lot of questions with low accuracy. We have decided to pay a fixed sum of 20 EUR for two hours. This is comparable to the salary of a student assistant in Germany.

Providers: We limit our study to 30 providers of web services. According to [1], these providers commit a wide range of privacy violations that are difficult to detect. Further, they have a significant market share in their respective domains, e.g., amazon for shops, google for search engines etc. We have excluded providers committing violations that are trivial to detect, e.g., 'no privacy policy'. We assign providers to participants automatically. We do so in a way that two participants have a minimal overlap of providers they review. Further, new providers are assigned whenever a user finishes a review. We have estimated that users review between 4 and 5 providers, i.e., this leads to approximately 10 ratings per provider with about 70 participants.

Violations: We model 6 categories of violations. The categories consider (1) the privacy policy, (2) data acquisition, (3) automated data processing, and how the providers handle (4) declarations of consent, (5) pseudonymous and (6) personalized profiles. Our taxonomy contains 43 questions and 31 violations.

Gold Standard: To verify that *CAPE* identifies violations correctly, we defined a Gold Standard. The same four experts that have defined the questions also define the correct answers for the 30 providers we consider in this study. We use these answers to measure the accuracy of the answers provided by 'regular' participants. The gold standard comprises 172 instances of the 31 violations. To fulfill the transparency requirement (R1), we treat similar violations independently. 'No information on cookie usage' and 'no information on the usage of web statistic tools' both violate information duties on automated data processing. To make the different causes for a violation transparent, we count them as two violations.

B. Study Procedure

We have structured our study in three phases. Each study group went through each phase.

Introduction (15 minutes): In the first phase we explain the concept and the user interface of *CAPE*.

Provider Review (100 minutes): In this phase the participants answer the questions provided by *CAPE*. Each participant reviews a different set of providers.

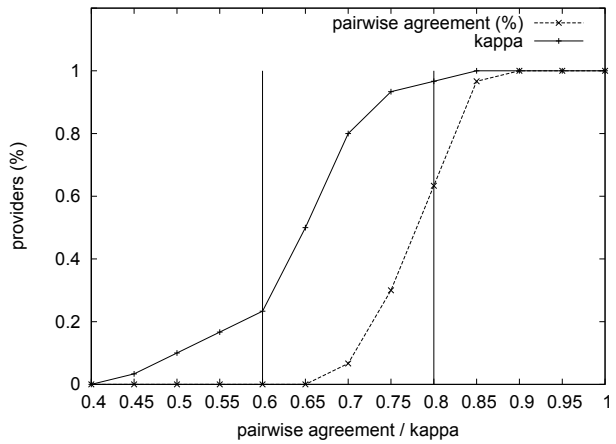

Figure 2. Degree of agreement

Completion (5 minutes): In the final phase each participant obtains a report showing the violations she has identified. Furthermore, we hand out a questionnaire to collect demographic data and to ask control questions.

C. Evaluation

1) Time for reviewing providers: Altogether, our 77 participants have answered 11,138 questions, i.e., around 145 questions per user. For 3.8% of the questions, the participants stated "don't know", i.e., for 10,711 questions they have given clear answers. On average, they have completely answered the set of questions for 4.18 providers.

CAPE asks different questions depending on answers given before. On average, each participant has answered 31.8 questions per provider, and each review has taken him approximately 22 minutes . This is in line with another study [11] where the participants needed between 18 and 26 minutes to read a privacy policy and to answer 5 questions.

Most participants have completely reviewed exactly 4 providers. We have measured the learning effect for these participants. For the first provider participants required 26 minutes on average. For both the third and forth provider the average time required is 21 minutes, i.e., an improvement of $\approx 20\%$. Comparing the answers to the Gold Standard (Section IV-C3), we have measured a constant number of correct and wrong answers.

Thus, the overhead *CAPE* raises, i.e., reading and answering the questions, does not lead to a significant increase in the time users need to review a provider (Q1), and people quickly become familiar with our approach.

2) Degree of Agreement: The concordance or inter-rater reliability, i.e., the degree of agreement among the users, is an important evaluation criterion for us. The more agreement between the users, the less controversial are the questions. We use two measures for this agreement.

$$\text{agreement(u,s,q)} = \frac{\text{\# answers to q identical to answers of u}}{\text{\# of all answers for q}} \quad (1)$$

$$\text{agreement}(s) = \frac{\sum_u \sum_q \text{agreement(u,s,q)}}{\text{\# users\_reviewing\_s} \cdot \text{\# questions}} \quad (2)$$

First, we compute the level of agreement among the users for a service provider s. We do so for each user u and provider s. The agreement is the fraction of answers to question q of other users identical to the answer of u (Equation (1)). Then we generalize this for all users and questions. This means that we compute the average of the pairwise agreement between all users having reviewed provider s (Equation (2)).

Second, we use Fleiss' Kappa [12]. It is a measure for the degree of agreement among several raters, here the participants, rating several objects, i.e., the questions. According to [13], kappa values < 0 mean poor agreement, values between 0 and 0.2 slight agreement, between 0.2 and 0.4 fair, between 0.4 and 0.6 moderate, between 0.6 and 0.8 substantial, and up to 1 perfect agreement.

Figure 2 shows the cumulative distribution function of the results for both the pairwise agreement and the kappa value for each provider. The pairwise agreement in percent and the kappa values are both represented by the horizontal axis. The vertical lines in the figure represent the borders for 0.4 to 1 for the quality of a kappa value described above.

Analyzing the kappa values, we see that for only approximately 20% of the providers the degree of agreement is moderate, i.e., kappa values in the left third of the diagram. For 80% of the providers in turn the kappa values achieved stand for a substantial to perfect agreement of the user answers. The pairwise agreement confirms this. It gives an intuitive measure for the clarity of the answers of most users. For 70% of the providers the agreement is above $\frac{3}{4}$. This is given by the intersection of a horizontal line through 0.3 on the vertical axis and the pairwise agreement function. Thus, answers have a high degree of agreement (Q2), and there are only few variations in understanding our questions and answering them for the different providers.

3) Correctness of answers (Gold Standard): Even though all participants may have given the same answer to a question, the answer can differ from the answer intended by the experts who defined the question. We now evaluate if the answers of our participants match the Gold Standard. V_{gold} denotes violations according to the Gold Standard, V_{usr} violations according to the user answers.

We first use a binomial test for the dichotomous variable {correct answer, wrong answer}. We test if we can reject the null hypothesis H_0 ='random agreement between the answers of the experts and the user'. We compute the test for the answers of each participant. Further, we do so for the consensus, i.e., the answer the majority of users has given for a particular provider and question. Second, we evaluate

for the majority answer to each question and provider the number of correctly identified violations ($\text{match} := V_{gold} \cap V_{usr}$), how many violations participants have not identified ($\text{misses} := V_{gold} \backslash V_{usr}$), and how many they have identified erroneously ($\text{false positives} := V_{usr} \backslash V_{gold}$).

Our participants have answered 9,050 (85%) questions correctly, i.e., in line with the Gold Standard. 1,661 (15%) answers do not match the Gold Standard. We discuss these numbers below. The binomial test confirms to reject H_0 on a significance level of 0.01, i.e., our approach allows people to answer our questions as intended by the experts. Further, we can reject H_0 with the same significance level of 0.01 when considering the majority answer of the users. To give an intuition for the quality of the answers we compute Cohen's kappa [14]. Cohen's kappa compares two ratings, here the participant answer and the gold answer, to multiple objects, i.e., the questions. Again using the scale from [13], we have measured a substantial to perfect agreement between the user answers and the Gold Standard for 79% of the users. Except for three outlier users, all others have a moderate agreement.

According to our Gold Standard, there are 172 violations. Using the majority answer, participants identified 177 violations. The overlap is 140 violations, i.e., 81%. The error is two sided: There are 19% misses and 20% false positives.

We conclude that the highly significant agreement between the Gold Standard and participants, as well as the large number of correctly identified violations are promising. It allows non-experts to detect violations (Q3). Nevertheless, the number of misses and false positives indicate that (some of) our questions can be misinterpreted. We have investigated this issue after having carried out the user study. We have observed that three of the 43 questions have led to most errors. As an example for falsely identified violations, our experts have interpreted the term 'we use cookies for automated login' in the privacy policy as a declaration of a persistent cookie. Some participants have not seen this as a clear hint for persistent cookies. Considering the misses, participants have mixed up the right to revoke consent and the right to opt-out, e.g., for pseudonymized user profiles. If those questions had been answered correctly, we would have obtained 88% matches, 13% misses and 14% false positives.

4) Comparison of social groups: In this section we analyze if the time needed to answer the questions, the degree of agreement, and the violations identified are different between the groups. We expect the law students to have the highest degree of agreement and correctly answered questions, and the pupils to have the lowest degree. We deem our results promising if the degree of agreement and the number of violations identified is high for all groups.

The groups have not varied much regarding the mean number of providers reviewed (cs 4.12, law 4.20, pup 4.38). On average, pupils have answered 130 questions, cs students and law students 148 questions. Pupils needed more time to read the privacy policies, but also to read the detailed explanation we had offered with each question. We find the effort required by each group comparable.

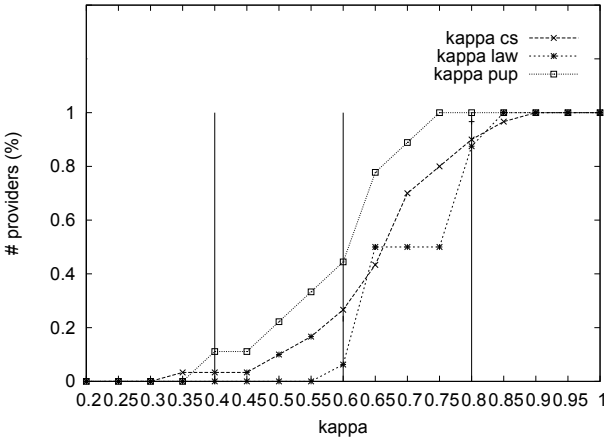

Figure 3. Agreement per social group

Table I
MATCHES, MISSES, FALSE POSITIVES

	Matches	Misses	FP
cs	0.81	0.19	0.28
law	0.80	0.20	0.20
pup	0.68	0.32	0.14

To calculate the degree of agreement between the users of each individual group we again use the Fleiss' kappa. The cumulative distribution functions of the kappa values are given in Figure 3, each curve represents one group. According to our expectation, pupils have the lowest degree of agreement, i.e., the curve of the pupils increases earlier than the the curves for the other groups. Law students have the highest degree of agreement. However, we have measured a substantial agreement for more than 50% of the providers (more than 90% for law students). For most of the remaining providers we still measured a moderate agreement. Again, note that values < 0 mean poor agreement and the chart starts with 0.2. Thus, the understanding of our questions and the reviewing of the providers is promising regarding all of the different groups.

Comparing the results of the individual groups to the Gold Standard (Table I) shows that the technical students and the law students behave very similarly. The violations identified by pupils are approximately 12% less than from the other groups. Interestingly, pupils have the lowest number of falsely identified violations. Our interpretation is that pupils, due to their lower experience in data protection, have read the full description and examples we have offered per question, and thus have answered our questions more accurately. The lower number of questions pupils have answered during the experiment time supports this. Further, the misses are due to a high number of "don't know" answers by pupils.

Summing up, our results show that all groups investigated can identify a wide range of violations. The knowledge on data protection or law required plays a minor role (Q4).

D. Discussion

In this section we discuss our approach and provide some further arguments why our study setup has been meaningful.

False Answers: Privacy practices not in line with data-protection law affect the reputation of a company. There might exist few companies trying to attack competitors, e.g., by giving erroneous reviews. However, attacks are beyond the scope of this study. Detailed surveys on potential attacks and countermeasures can be found in [15], [16].

Majority Vote: Majority vote is the simplest mechanism to come to a consensus. An alternative scenario would be to weight users higher when data-protection authorities confirm a violation detected, or to apply mechanisms related to [17]. We have used majority vote for two reasons: First, we did not want to reveal our Gold Standard during the study. Second, data protection has many aspects, and we expect participants not to have the same answer quality in all fields (understanding the potential of personalized profiles, automated data processing, giving consent, etc.) Thus, to weight the answer quality of a participant for each aspect of data protection, the 4 reviews a participant has generated on average are not sufficient from our perspective.

Perspective: For the study, participants have identified violations from an external perspective. However, by extending the taxonomy, our approach can easily be adapted to, e.g., a company. Then the employees form the community, probably with only partial knowledge of their department, answer the question they can, and the individuals responsible for data protection within the company verify if the answers comply with what the customers gave their consent to.

V. Conclusions

Data protection and abidance by the law is an important issue for online services. However, previous studies have shown that providers frequently do not conform to law, and authorities do not control them efficiently. So far, without legal expertise, Internet users do not have a chance to identify violations.

In this work we have proposed an approach to collaboratively identify data-protection violations. We map legal expertise to a taxonomy of intuitive questions. By extensive user studies, we have shown that non-experts can efficiently identify privacy violations with high statistical significance. They have identified 81% of those violations experts find, with only a small variance regarding the different social groups investigated.

We expect the social impact of our results to be high: Companies can evaluate themselves using our approach, Internet users can compare providers based on uniform assessment criteria, and data-protection commissioners can focus on violations identified with a high statistical significance. Thus, intuitive approaches like *CAPE* might help tackling compliance deficits of data-protection law.

Acknowledgments

This work was partly funded by DFG BO2129/8-1. We thank Martin Helfer for his intensive support.

References

[1] T. Burghardt *et al.*, "A study on the lack of enforcement of data protection acts," in *e-Democracy, Greece*, 2009.

[2] S. Zhang, J. Ford, and F. Makedon, "A privacy-preserving collaborative filtering scheme with two-way communication," in *ACM Conference on Electronic Commerce*, 2006.

[3] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *WPES*. ACM, 2005.

[4] G. Yee and L. Korba, "Comparing and matching privacy policies using community consensus," in *IRMA*, 2005.

[5] C. Heidinger, E. Buchmann, and K. Böhm, "Collaborative data privacy for the web," in *EDBT Workshops*. ACM, 2010.

[6] N. H. Vyasa *et al.*, "Towards automatic privacy management in Web 2.0 with semantic analysis on annotations," Rutgers University, Tech. Rep., 2009.

[7] M. Marchiori, "The platform for privacy preferences 1.0 (p3p1.0) specification," W3C, 2002.

[8] P. Beatty *et al.*, "P3P adoption on e-commerce web sites: A survey and analysis," *IEEE Internet Computing*, 2007.

[9] C. A. Brodie, C.-M. Karat, and J. Karat, "An empirical study of natural language parsing of privacy policy rules using the sparcle policy workbench," in *SOUPS*. ACM, 2006.

[10] K. Engisch, *Introduction to legal thinking (de)*, 10th ed., T. Würtenberger, Ed. Kohlhammer, 2005.

[11] A. M. McDonald and L. F. Cranor, "The cost of reading privacy policies," *A Journal of Law and Policy for the Information Society*, 2008.

[12] F. JL., "Measuring nominal scale agreement among many raters ." in *Psychol Bull*, 1971.

[13] J. Landis and G. Koch, "The measurement of observer agreement for categorical data." *Biometrics*, 1977.

[14] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, 1960.

[15] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decis. Support Syst.*, 2007.

[16] K. Hoffman, D. Zage, and C. Nita-Rotaru, "A survey of attack and defense techniques for reputation systems." *ACM Comput. Surv.*, 2009.

[17] D. Prelec, "A Bayesian truth serum for subjective data," *Science*, vol. 306, no. 5695, p. 462, 2004.