ÇORBA: crowdsourcing to obtain requirements from regulations and breaches

Hui Guo[1] · Özgür Kafalı[2] · Anne-Liz Jeukeng[3] · Laurie Williams[1] · Munindar P. Singh[1]

Abstract

Context Modern software systems are deployed in sociotechnical settings, combining social entities (humans and organizations) with technical entities (software and devices). In such settings, on top of technical controls that implement security features of software, regulations specify how users should behave in security-critical situations. No matter how carefully the software is designed and how well regulations are enforced, such systems are subject to breaches due to social (user misuse) and technical (vulnerabilities in software) factors. Breach reports, often legally mandated, describe what went wrong during a breach and how the breach was remedied. However, breach reports are not formally investigated in current practice, leading to valuable lessons being lost regarding past failures.

Objective Our research aim is to aid security analysts and software developers in obtaining a set of legal, security, and privacy requirements, by developing a crowdsourcing methodology to extract knowledge from regulations and breach reports.

Method We present ÇORBA, a methodology that leverages human intelligence via crowdsourcing, and extracts requirements from textual artifacts in the form of regulatory norms. We evaluate ÇORBA on the US healthcare regulations from the Health Insurance Portability and Accountability Act (HIPAA) and breach reports published by the US Department of Health and Human Services (HHS). Following this methodology, we have conducted a pilot and a final study on the Amazon Mechanical Turk crowdsourcing platform.

Results ÇORBA yields high quality responses from crowd workers, which we analyze to identify requirements for the purpose of complementing HIPAA regulations. We publish a curated dataset of the worker responses and identified requirements.

Conclusions The results show that the instructions and question formats presented to the crowd workers significantly affect the response quality regarding the identification of requirements. We have observed significant improvement from the pilot to the final study by revising the instructions and question formats. Other factors, such as worker types, breach types, or length of reports,

✉ Hui Guo
  hguo5@ncsu.edu

Extended author information available on the last page of the article.

do not have notable effect on the workers' performance. Moreover, we discuss other potential improvements such as breach report restructuring and text highlighting with automated methods.

Keywords Regulatory norms · Sociotechnical systems · HIPAA

1 Introduction

The development of sociotechnical systems requires the developers to comply with existing regulations that describe the expected behaviors of software and its users, particularly in domains that deal with private user information. Existing studies (Breaux and Antón 2008; Ghanavati et al. 2014; Hashmi 2015; Maxwell and Anton 2009; Siena et al. 2012) model or extract information from such regulatory documents to help with the elicitation of requirements for developers or compliance checking for legal purposes.

One of the most studied regulations is the Health Insurance Portability and Accountability Act of 1996 (HIPAA) (HHS 2003) in the healthcare domain, which is a legislation on data privacy and security regarding medical information. Textbox 1 shows an example clause from HIPAA regarding devices and media containing electronic protected health information (PHI).

Textbox 1

§164.310 - Physical safeguards
A covered entity or business associate must, in accordance with §164.306:

§164.310(d)(1) - Standard: Device and media controls.
Implement policies and procedures that govern the receipt and removal of hardware and electronic media that contain electronic protected health information into and out of a facility, and the movement of these items within the facility.

This clause imposes a requirement on a covered entity (CE) or a business associate (BA) regarding hardware and media.

Regulatory text is often abstruse and unclear as to requirements. Breaches are frequently caused by social (user misbehavior) and technical (flaws in software) violations. Breach reports (HHS Breach Portal 2016; Murukannaiah et al. 2017; Verizon 2016), often legally mandated, describe cases where deployed systems fail, or are maliciously or accidentally misused (Matulevičius et al. 2008; Sindre and Opdahl 2005), and suggest actions to prevent, detect, and recover from future breaches (Liu et al. 2015; Riaz et al. 2016). Breach reports can help security analysts understand regulations by providing instances where regulations are violated. For the above clause, breaches can help in understanding why and how the "policies and procedures" should be implemented.

In recent years, healthcare data breaches, caused by outside attacks as well as insider misconducts, have brought HIPAA increasing prominence. US law now requires the US Department of Health and Human Services (HHS) to post each breach of unsecured PHI affecting 500 or more individuals (HHS Breach Portal 2016). Textbox 2 provides an example report of the violation of the above clause. This report includes actions taken by the responsible party after the breach, which could inform prevention of, or recovery from, similar breaches. Such reports sometimes include actions that other parties, such as Office for Civil Rights (OCR), took after the breaches.

Textbox 2

1. An unencrypted portable data drive was lost by a pharmacy resident of the Arnold Palmer Hospital, a part of the covered entity (CE).
2. The drive contained the protected health information (PHI) of 586 individuals, including names, birth weights, gestational age, admission and discharge dates, medical record numbers, and some transfer dates.
3. The missing drive also stored personal items, a research study proposal, and two spreadsheets containing limited information on 586 babies who were part of a study.
4. The CE provided breach notification to HHS, the media, and to the parents of the affected individuals because they were all minors.
5. Substitute notice was posted on the CE's website.
6. The CE updated its policies and procedures for its data loss prevention system and added controls.
7. The CE retrained the resident involved in the loss of data and provided additional information to all employees and medical staff members regarding the use of portable data devices through education and published articles.
8. OCR obtained assurances that the CE implemented the corrective actions listed above.

We adopt norms (Barth et al. 2006; Hao et al. 2016; Kafalı et al. 2017; Singh 2013; Von Wright 1999) to formalize regulations and breaches (as violations of norms). Norms (here, deontic norms including commitments, authorizations, and prohibitions) provide a compact, yet expressive formalization. However, extracting norms from text is nontrivial: natural language processing (NLP) methods yield low accuracy (Dam et al. 2015; Gao and Singh 2014; Kashyap et al. 2016; Hashmi 2015), and trained experts are costly (Breaux and Schaub 2014). Like all textual artifacts, breach reports are often ambiguous, inconsistent, and incomplete, which makes it hard to extract useful information, thereby losing valuable knowledge. *Crowdsourcing* approaches (Breaux and Schaub 2014; Dean et al. 2015; Getman and Karasiuk 2014; MacLean and Heer 2013; Murukannaiah et al. 2016; Reidenberg et al. 2015; Wilson et al. 2016) that rely on collective human intelligence have gained increasing attention for information extraction tasks regarding legal text. However, such approaches have not been applied to "end user reported" artifacts, such as breach reports, or with the aim of connecting multiple artifacts. Accordingly, we present ÇORBA,[1] a methodology that leverages human intelligence via crowdsourcing to obtain requirements by extracting and connecting key elements from regulations and breach reports. We represent the obtained requirements as a set of regulatory norms to provide a structured yet compact presentation for practitioners. Specifically, we address the following core research question toward the obtaining of requirements in the form of norms from these textual artifacts:

RQ: How can we design a crowdsourcing task to effectively extract security requirements from regulations and breach reports as norms, and what factors affect the performance of crowd workers for this task?

---

[1]ÇORBA is Turkish for "Soup": It acts as a memorable name for our methodology (close to an acronym), and reflects the mixture of multiple artifacts contained in our study.

This research question contributes toward security requirements engineering through (i) the clarification of ambiguous requirements, and (ii) the elicitation of previously unknown security requirements. In previous work, Kafalı et al. (2017) formally investigated the connections between the knowledge contained in regulations and breach reports via normative reasoning. Here, we empirically validate such connections by presenting our crowd workers with a selected set of regulations from HIPAA that are associated with the breach reports from HHS. We answer this research question by investigating the effect of the variables in our studies, and gathering insights into the application of our methodology.

Additionally, we show that more concise and more structured breach reports can lead to more accurate extraction of relevant actions regarding the avoidance and prevention of future breaches. We also discuss how automated methods can help trim a breach report by identifying key sentences.

Our contributions include (i) a crowdsourcing methodology for extracting normative elements from regulations and breach reports as well as its empirical evaluation; (ii) demonstration of the need for concise and structured reporting of breaches; and (iii) a curated dataset of the results of applying ÇORBA on a selected set of regulations from HIPAA and breach reports from HHS, including crowd worker responses and the extracted norms.

The rest of the paper is structured as follows. Section 2 reviews the necessary background. Section 3 describes our crowdsourcing methodology. Section 4 details how we deploy our methodology on mTurk. Section 5 presents our results. Section 6 presents the key findings. Section 7 discusses the threats to validity as well as additional findings. Section 8 concludes with future directions.

2 Background and Related Work

Norms We now describe our formal representation for norms. A *norm* in the particular sense we adopt here is a directed relationship between a *subject* (the party on whom the norm is focused) and an *object* (the party with respect to whom the norm arises) that regulates their interactions (Singh 2013). Each norm also specifies an *antecedent*, the conditions under which the norm is effective, and a *consequence*, the conditions that fully satisfy the norm. A set of norms describes the social architecture of a sociotechnical system. We consider three types of norms: commitments (c), authorizations (a), and prohibitions (p). A *commitment* means that its subject is committed to its object to bringing about the consequent if the antecedent holds. An *authorization* means that its subject is authorized by its object to bring about the consequent if the antecedent holds. A *prohibition* means that its subject is prohibited by its object from bringing about the consequent if the antecedent holds.

We write a norm as $n(\text{SBJ}, \text{OBJ}, \text{ant}, \text{con})$, where n, its type, is one of $\{c, a, p\}$; SBJ is its subject; OBJ its object; ant its antecedent; and con its consequent. This conception of norms is compatible with previous models (Barth et al. 2006; Singh 2013) and avoids some of the shortcomings of traditional deontic logic concepts pointed out by Von Wright (1999).

Norms provide a natural formal representation for security and privacy requirements (Kafalı et al. 2016a, b). In addition to the required actions, desired states and their conditions, norms describe the direction of accountability. The clause in Textbox 1 may be formalized as this:

> Type: *c*: Commitment
> Subject: COVERED ENTITY or BUSINESS ASSOCIATE
> Object: HHS
> Antecedent: hardware and electronic media contain electronic PHI.
> Consequent: implement policies and procedures that govern receipt, removal, and movement of hardware and electronic media

The following norms are the extracted norms from Textbox 2 during our experiments of ÇORBA.

> Type: *p*: Prohibition
> Subject: EMPLOYEE (Sentence 1)
> Object: COVERED ENTITY (Sentence 1)
> Antecedent: portable devices contain PHI (Sentence 1)
> Consequent: lose portable devices (Sentence 1)

> Type: *c*: Commitment
> Subject: COVERED ENTITY (Sentence 7)
> Object: PATIENTS (Sentence 2)
> Antecedent: TRUE (at all times)
> Consequent: train employees on data loss, data protection (Sentence 7)

The second norm is actionable and connected to the above HIPAA clause and can help prevent similar breaches. We omit other norms mentioned in this report for brevity. This negative example, i.e., the breach of information in Textbox 2 provides security analysts a clearer understanding as to how the clause in Textbox 1 can be interpreted. Moreover, the breach report provides information regarding how to mitigate the breach, namely a data loss prevention system and improved training of employees.

Studies have been conducted on norm extraction from business contracts (Gao and Singh 2014) and regulatory documents (Hashmi 2015) where normative relations are explicitly described. Social norms or conventions in a community, such as open source software repositories, can be implicit and therefore are difficult to mine (Dam et al. 2015; Savarimuthu and Dam 2014). Breach reports describe norms in an implicit or negative way. Albeit fruitful, norm extraction from these textual artifacts poses new challenges.

Security Related Artifacts Analyzing breaches helps analysts understand how failures, e.g., unintentional or malicious actions by the software or its users, affect compliance with applicable regulations. Gürses et al. (2008) develop heuristics for designing usable privacy interfaces in online social networks based on investigation of privacy breaches. Their analysis helps in the understanding of privacy conflicts and trade-offs revealed by such breaches, with respect to each stakeholder's viewpoint. Kafalı et al.'s (2017) framework compares what happened in a breach with what the regulation states. However, they do not provide a way of extracting norms from text.

Crowdsourcing Whereas security requirements can be extracted through the analysis of policies and regulations, analyzing such natural language text is labor intensive and tedious for analysts. Crowdsourcing (Breaux and Schaub 2014; Dean et al. 2015; Getman and Karasiuk 2014; MacLean and Heer 2013; Patwardhan et al. 2018; Reidenberg et al. 2015; Wilson

et al. 2016) of information extraction from legal text is a promising and popular approach to address this challenge. Breaux and Schaub (2014) propose experiments to compare the effectiveness (accuracy and cost) of untrained crowd workers on a requirements extraction task with the effectiveness of trained experts (i.e., requirements engineers). Their task includes extracting data collection, sharing, and usage requirements from privacy policies. Breaux and Schaub report that they could reduce manual extraction cost by up to 60% for some policies while preserving task accuracy, and for some policies increase accuracy by 16%, based on their ways of task decomposition. They continue using crowdsourcing, combined with NLP techniques, to extract privacy goals (Bhatia et al. 2016). Reidenberg et al. (2015) investigate how privacy policies are perceived by expert, knowledgeable, and typical users, and did not find significant differences among them.

Text Analysis The task of extracting useful information in a formal representation from textual documents, such as security-related textual artifacts, is of great importance. Researchers start from designing and proposing systematic methodologies for manual extraction. Breaux and Antón (2008) have developed a methodology for manually extracting formal descriptions of rules, such as rights and obligations, that govern information systems from regulatory texts. They represent results from a case study on the text of HIPAA Privacy Rule. Hashmi (2015) presents a methodology for the extraction of legal norms from regulatory documents that emphasizes logical structures for reasoning and modeling to facilitate compliance checking. Systematic manual extraction methodologies are helpful for domain experts to analyze text, but may not be applicable to nonspecialists, such as typical workers in crowdsourcing projects. Also, the transition from a manual extraction process to an automated one is not straightforward and needs further investigation.

Automatically converting textual artifacts into a formal representation is challenging, and may involve semantic comparison, summarization, and rephrasing. Riaz et al. (2014) describe a tool-assisted process that incorporates machine learning for identifying security requirements from text. They empirically derive a set of context-specific templates to translate such objectives and goals into security requirements. Slankas and Williams (2013) propose an automated process for extracting access control policies implicitly and explicitly defined in natural language project artifacts. Zeni et al. propose the NómosT tool (2018) to help users construct goal-based representation of legal requirements semi-automatically by identifying and using metadata in legal texts based on their Nómos framework (Siena et al. 2012) and GaiusT framework (Zeni et al. 2015, 2017). Sleimi et al. (2018) propose automated extraction rules for semantic metadata based on NLP, which can help with understanding legal provisions. Current such automated methods for extracting requirements from text either require domain-specific knowledge and heuristics and therefore are costly to migrate to other domains, or do not perform end-to-end extraction. We believe that using crowdsourcing for the extraction task is more generalizable, but automated methods can be leveraged to facilitate the extraction, which we discuss in Section 6.2.1.

3 Çᴏʀʙᴀ Crowdsourcing Methodology

Çᴏʀʙᴀ, our crowdsourcing methodology for extracting norms from text, consists of data collection and evaluation steps. We design a survey for data collection, which consists of two parts: (Part I) reading through a tutorial to help crowd workers understand what is expected from them; and (Part II) answering multiple choice and free text response questions. Each crowd worker performs Part II for five breach reports. In a pre-survey, we ask workers

whether they have seen legal text before to learn about their experience with the task. At the end of the survey, workers rate the difficulty of the survey tasks. Details of each task are described in Section 3.1. In the evaluation step, we analyze the surveys received from workers by evaluating their responses according to qualitative scales. Details of how we develop the qualitative scales and evaluate worker responses are described in Section 3.2.

Our core research question investigates the factors to consider when applying crowd-sourcing to the extraction of norms from regulations and breach reports. Specifically, we experiment on ÇORBA targeting the following refined questions on factors that affect quality:

RQ$_1$: How do the properties of a breach report, including its length, complexity, perceived difficulty, and type of breach, affect the crowd workers' performance?

RQ$_2$: Are some elements of a norm more challenging than others to extract from regulations and breach reports?

Figure 1 describes how we conducted our study. We first investigated how project settings and task descriptions affect the quality of worker responses in a pilot study. According to the preliminary analysis, we revised the settings and tasks for the final data collection.

3.1 Tasks

We create our survey with tasks that aim to extract key elements from breach reports and regulations as well as to understand any relations between them. Each task includes multiple choice questions or questions that require free text input. Each survey starts with a breach
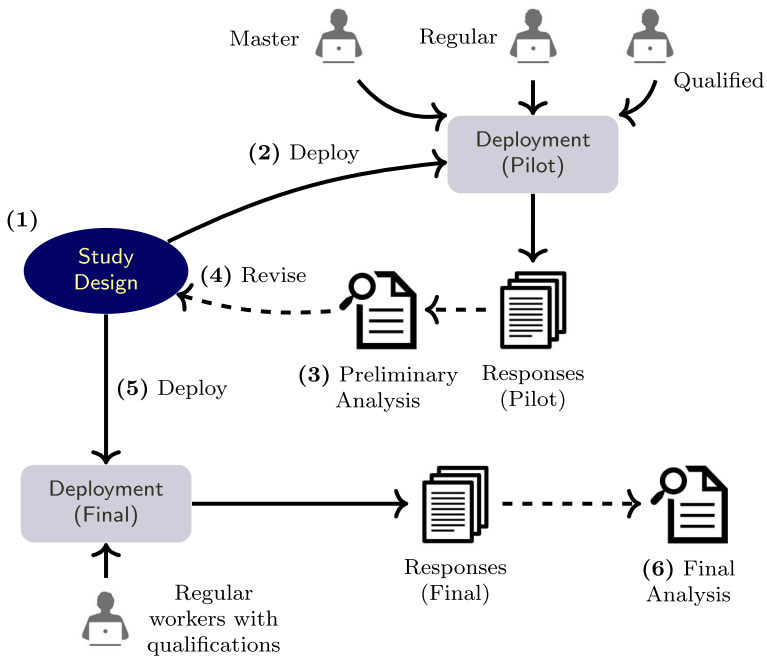

Fig. 1 Overview of our study to evaluate ÇORBA. Numbers represent various steps. Solid arrows are data collection steps. Dashed arrows represent evaluation steps

report broken down into separate sentences. We include sample screenshots from the survey in the Appendix (see Figs. 7–8 for the tutorial and Fig. 9 for a worker response). More study materials can be found here: https://goo.gl/xda2nQ. The workers perform the following four tasks regarding each breach:

Malice: Task 1 is to understand whether a breach occurred due to malice or an accident (human error). We ask two five-point Likert scale (Allen and Seaman 2007) ("Strongly disagree," "Disagree," "Neutral," "Agree," "Strongly agree") questions: "The incident described in the above breach report is caused by malicious intent." and "The incident described in the above breach report is due to accidental human error."

Breach: Task 2 is to extract normative elements from the breach report. Since the general worker population might not be familiar with the concept, we created simplified questions to extract such elements. In particular, workers answer these questions and identify relevant sentences from the breach report:

(Breach.Action: Task 2.1)   consequent: "What is the necessary action to prevent this breach?" The answer to this question would be the consequent of the norm that would mitigate this breach.

(Breach.Who: Task 2.2)   subject: "Who should have taken that action?" The answer to this question would be the subject of the norm.

(Breach.Condition: Task 2.3)   antecedent: "In what circumstances should the action be taken?" The answer to this question would be the antecedent of the norm that describes the conditions under which the norm is in effect.

(Breach.Whom: Task 2.4)   object: "Who (or what organization) is affected by the breach?" The answer to this question would be the object of the norm.

Relevance: Task 3 is to understand whether the HIPAA clause covers what happened in the given breach. We ask workers to rate a list of provided HIPAA clauses on a four-point Likert scale: from "not relevant" to "exactly the case in the breach." In our study, we chose the targeted clauses before the deployment semi-randomly, as described in Section 4.2.

Regulation: Task 4 is to extract normative elements from a regulation clause. We show workers the most relevant clause (according to us), and ask the following questions:

(Regulation.Who: Task 4.1)   subject: "Who is the responsible party/individual for the above policy?"

(Regulation.Whom: Task 4.2)   object: "To whom are they responsible?"

(Regulation.Action: Task 4.3)   consequent: "What action should the responsible party (not) take to comply with the policy?"

(Regulation.Condition: Task 4.4)   antecedent: "In what conditions should they take the above action?"

We show Task 4 to a worker on a separate survey page after the worker completes the first three tasks, since we do not want to bias the worker's answers for Task Relevance (Task 3).

We require workers to read through a tutorial before beginning the survey. The tutorial includes a typical breach report, the task questions, and our preferred answers to them. Additionally, we provide compact and intuitive explanations to the given answers. After the survey, we present workers with two additional questions regarding the difficulties of understanding the breach and the regulation clause, respectively.

3.2 Evaluating Worker Responses

The first and third authors act as evaluators to assess the quality of the worker responses after each phase. The evaluators have been trained on regulations, norms, and text analysis. For Likert scale questions, i.e., Tasks Malice (Task 1) and Relevance (Task 3) in Section 3.1, the evaluators judge the quality of the response by reading a worker's response to multiple questions. The workers' answers can be close to what the evaluators expect, e.g., compatible, or far from what the evaluators expect, e.g., contradicting. Specifically, we use the following scale:

(D)   Deficient The worker's response is empty, contradicting, or far from the expected answer. A contradicting response can be inferred from multiple answers that contradict with each other. Answers where a worker responds randomly, or with a discernible pattern (e.g., the same answer to every single question), are also considered as deficient.

(F)   Fair. The worker's response is fair if it contains answers with varying quality. The response can also contain too many "neutral" or "somewhat" choices where the answers are clear.

(C)   Compatible. The worker's response is within an acceptable range from the expected answer. If the evaluator thinks the answer should be "Strongly agree," a response of "Strongly agree" or "Agree" is compatible.

For free text responses (Tasks Breach and Regulation in Section 3.1) from workers, we use the following scale:

(B)   Bad. The response is empty, random, or not relevant to the task. For example, some workers copy and paste random sentences from the breach reports to fill the survey.

(U)   Unclear. The response is too long, not easily understandable, or not logically sound. For example, one worker answered "no longer employed" as the condition for "Federal Law Enforcement" to "add safeguard."

(N)   Not on topic. The response is somewhat relevant, but not to the point. For example, some workers identify recovery actions "Return the stolen documents" or "Offer free credit monitoring" as "the necessary action to prevent this breach." These answers are not irrelevant, but they are not the specific actions to prevent the particular breach.

(A)   Acceptable. The response is relevant to the task, but it could be phrased better. For example, some workers copied phrases from the breach report, and used improper tense.

(E)   Excellent. The response is composed of one or more of the potential answers. For example, workers usually answer with only one of "Covered Entity," "Patients," or "HHS" to the question "Who (or what organization) is affected by the breach?"

The evaluators first individually evaluate each worker response. They then discuss and resolve any disagreements that they might have. For the free text responses, we consider disagreements between a relatively worse response (B, U, N) and a better response (A, E). For multiple choice responses, we consider disagreements between all evaluations. Note that we convert the evaluations to numeric values for performing statistical analysis: Grades (D, F, C) are transformed to (0.0, 0.5, 1.0) respectively; and Grades (B, U, N, A, E) are transformed to (0.0, 0.25, 0.5, 0.75, 1.0) respectively. The scores of a worker's answers to one breach report are averaged to form one score for the quality of the combined response.

Based on the evaluations, the evaluators work together to combine, clean, and organize the worker responses, in order to obtain structured norms for each breach report, which are the final results of applying this methodology.

4 Amazon Mechanical Turk Study

We applied ÇORBA on Amazon Mechanical Turk (mTurk). Our study design was approved by NCSU's institutional review board (IRB). Each task that a worker performs in mTurk is called a human intelligence task (HIT). A batch is a group of HITs with similar settings, which a requester can start and finish at the same time. In our project design, workers needed to sign an *informed consent* form before they accepted the HIT. We instructed each worker to participate in our study no more than once, since we posted several batches throughout our study.

4.1 Worker Groups

MTurk Requesters can create HITs with different Qualification requirements,[2] and only workers that meet these Qualifications can access the HITs. In our study, each batch required a Master Qualification, no Qualification, or a list of pre-defined Qualifications. Accordingly, we formed the following three groups of workers with the corresponding Qualifications.

Master   workers are described by Amazon as "elite groups of workers who have demonstrated accuracy on specific types of HITs on the Mechanical Turk marketplace." Amazon maintains the standard and statistics of Master workers, and charges more fees for projects that requires a Masters Qualification, which only Master workers can access.

Regular   workers are workers who do not need to meet any Qualifications.

Qualified   workers are regular workers with additional Qualifications, specifically, those who (i) have 95% or higher approval rate from previous HITs, (ii) have completed at least 100 HITs, and (iii) are located in an English-speaking country.

4.2 Study Procedure

After some initial investigation of the HHS breach reports dataset and extensive discussion, we selected 39 breach reports that reflected the distribution of breach types. In a previous study (Kafalı et al. 2017), we have hired trained graduate students to identify the most relevant HIPAA clauses to these breaches, by manually examining through every clause in the HIPAA security and privacy sections. The 39 breaches were mapped to eight clauses. We added 13 random clauses from the same sections, so we could present five clauses, one of which was the most relevant, to each crowd worker for Task Relevance (Task 3) of each breach. We selected one breach report and answered the questionnaire as a tutorial, leaving 38 breach reports for the workers. Once workers read the tutorial, they were shown links to five questionnaires regarding five randomly-chosen breaches out of 38. Once workers finished all five questionnaires, they were prompted a unique confirmation code, which could be submitted on mTurk to finish the HIT and get paid. All worker responses were recorded in our database, along with the time workers took to finish each breach questionnaire.

In our pilot study, we executed multiple batches with our initial task design involving different types of workers. Based on our findings, in our final data collection project, we adjusted the survey questions and instructions, and involved only qualified workers, which we discuss in detail in Section 5.1. In addition, we shortened five breach reports producing three distinct versions (see Section 5.2). These reports had received responses with the

---

[2]https://www.mturk.com/mturk/help

Table 1 mTurk deployment statistics

Study element	Pilot	Final
# survey questions (per breach)	15	16
# master workers	13	0
# regular workers with no qualifications	12	0
# regular workers with qualifications	13	42
# breach reports / # unique questionnaires	38	53
# regulation clauses for Task Relevance	21	21
# regulation clauses for Task Regulation	8	8

lowest quality. Thus, the total number of breaches used was 53 ($38 + 5 \times 3$). Statistics about the pilot and final studies, such as numbers of participating workers, are shown in Table 1. In the pilot study, we rejected workers that did not provide correct confirmation codes. In the final study, we examined all workers' responses as they arrived, and rejected workers that did not follow our instructions for quality.

4.3 Compensation

Before the deployments, we estimated the time for finishing a survey as 20 minutes based on preliminary test runs. In our pilot study, we compensated each accepted worker with $3 as base payment, and promised bonus for good results, according to suggested rate on mTurk (Downs et al. 2010). However, actual workers finished the survey (with five breach reports) in the pilot study with a median time of 32 minutes. We compensated workers with C, A, and E responses with the bonus payment up to $2 (later increased to $3, see Section 5.2). Even though ours was not a typical mTurk task, with a high variance of completion times, most workers reflected that the compensations were fair, with a few preferring more reward.

5 Crowdsourcing Study Results and Revisions

We compiled the assessments of our evaluators for the two studies: pilot and final. We first present the results from the pilot study. Based on the results, we describe our revisions. Then, we present the results from the final study.

5.1 Pilot Study Results

In the pilot study, we rejected workers who did not provide a correct confirmation code or whose answers were clearly incorrect. For example, one worker answered every question with a single period character. We rejected three workers (7.3%).

To compare the performances of workers on the tasks, we convert our qualitative scales to a quantitative (0–1) scale, as described in Section 3.2. We average the ratings of the two evaluators and then combine the ratings for all tasks to get a quantitative evaluation for the quality of a response.

RQ$_1$ Properties of a breach report include the length of the report, the perceived difficulty of the extraction, and the type of the breach. We do not find a significant correlation ($p = 0.544 > 0.10$) between the length of breach reports (number of sentences) and the
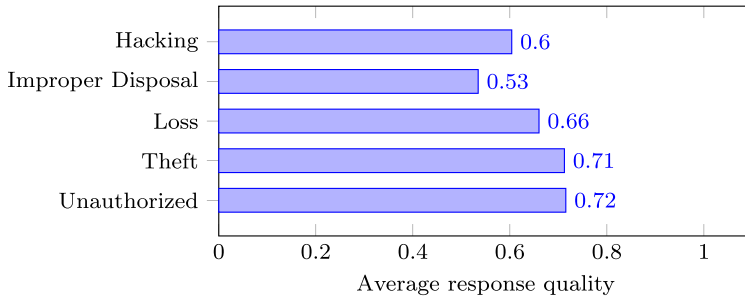
Fig. 2 Performance across breach types in the pilot

quality of worker responses. Note that we recorded workers' thoughts on the difficulty of each questionnaire. There was a significant ($p = 0.0008 < 0.01$) yet weak correlation between the perceived difficulty and number of sentences in breach reports with a correlation coefficient of 0.229. This association may not be practically relevant, however. We find larger variance in response quality among breach types, as provided by HHS. Figure 2 shows the average response quality for each type. For improper disposal cases, workers have provided less than optimal answers like "dispose of notes properly" or "retraining employees," which only apply in certain contexts. For hacking breaches, workers may have lacked necessary technical knowledge, and provided answers like "don't make server accessible to internet all the time."

RQ$_2$ We answer this research question by examining the response quality for each survey task, since most of them (Tasks 2 and 4) correspond to elements of the norms to be extracted. We find that the quality presented great variance across tasks. Figure 3 shows the average quality for each task. Answers to Tasks Breach.Condition (Task 2.3) and Regulation.Condition (Task 4.4) had the lowest quality. These questions are regarding the antecedent of a norm (as explained in Section 2). Many workers answered these questions
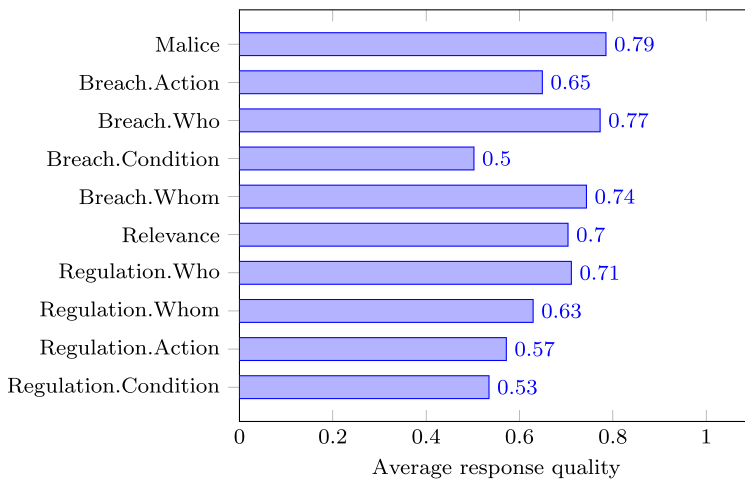

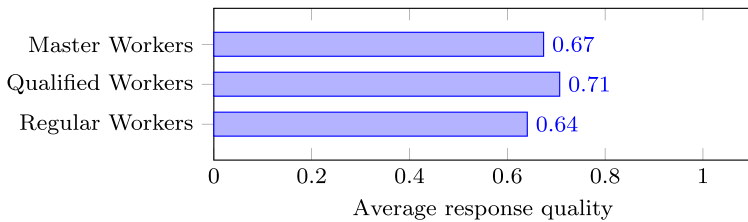**Fig. 3** Performance across tasks in the pilot

Fig. 4 Performance across worker groups in the pilot

with phrases or sentences that did not describe conditions or circumstances. Answers to Tasks Breach.Action (Task 2.1) and Regulation.Action (Task 4.3) also had lower quality than others. These questions are regarding the actions, or the consequences, of norms. Some workers tended to provide short and insufficient answers to these questions. We notice that a lot of workers copied phrases and sentences from reports, which could be acceptable answers, but usually corresponded to a recovery action rather than the cause of a breach (which is what Task Breach.Action asks), and were given in the wrong tense.

Worker Groups Figure 4 shows the average quality for each worker group. Our analysis does not suggest any statistically significant differences regarding the quality of responses among worker groups. In fact, qualified workers, i.e., regular workers that qualified for our specifications, yielded results with higher quality on average. This result, while surprising, is compatible with previous research on comparing experts with typical users for policy understanding (Reidenberg et al. 2015). One possible reason for this result is that, although master workers yield on average high quality results on a variety of work, they are not necessarily familiar with text analysis, especially given that our project is an unusual one. For example, some master workers may not be native English speakers.

Worker Effort Workers spent varied amounts of time on our survey. The average time workers spent on the tutorial was 12.7 minutes, and the median was two minutes. Average time spent on a questionnaire for one breach report was 10.3 minutes, and the median was six minutes. We have not found significant differences of the finish times among the worker groups (on average, 10.6, 10.0, and 10.1 minutes for master, regular, and qualified workers, respectively). On average, workers finished our survey (with five breach reports) in 64.2 minutes. Median time was 32 minutes.

5.2 Study Revisions After Pilot

Based on the preliminary analysis, we revised the following:

Survey Tasks According to the relative task difficulties identified in Section 5.1, we significantly revised the tasks that workers performed worse on, and slightly revised other tasks. Specifically, we revised the following tasks from Section 3.1:

(Breach.Action: Task 2.1)   consequent: "What is the necessary action to prevent this breach? Please start your answer with a verb. Do not use past tense." We limited this question to prevention actions, and added specific instructions to increase the quality of responses.

(Breach.Who: Task 2.2)   subject: "Who should have taken that action? Please keep in mind that this action is the same action you have given in the previous question." Based on the responses we analyzed, we included a reminder in this question.

(Breach.Condtion: Task 2.3)   antecedent: "Consider you are the responsible person for carrying out the action in Question 1. Do you always have to take that action, or only in specific circumstances? Please start your answer with 'when' or 'if'." We determined that the original Task Breach.Condition for this task was vaguely stated. Therefore, we connected it to Task Breach.Action, and gave workers specific directions for answering the question.

(Breach.Whom: Task 2.4)   object: "Who (or what organization) is affected by the breach? Please state ALL parties that (might) have been affected. These are the people your answer to the second question should have protected." We included some reminders in this question.

(Regulation.Who: Task 4.1)   subject: "Who is the responsible party/beneficiary for the above policy? State the name or role of this party/individual." After analyzing the worker responses, we identified that asking specific questions based on whether the regulation clause was an authorization or a commitment might increase the quality of responses. We kept "responsible party" for commitments and added "beneficiary" for authorizations.

(Regulation.Whom: Task 4.2)   object: "To whom are they responsible? / Who authorizes the beneficiary?" We revised this question to remind the worker of the distinction between an authorization and a commitment.

(Regulation.Action: Task 4.3)   consequent: "What action should/may the responsible party/beneficiary take to comply with the policy?" We revised this question to remind the worker about the distinction between an authorization and a commitment.

(Regulation.Condition: Task 4.4)   antecedent: "Consider you are the responsible person for carrying out the action in Question 3. Do you always have to /are you always authorized to take that action, or only in specific circumstances? Please start your answer with 'when' or 'if'." we connected this question to Task Regulation.Action, and gave workers specific directions for answering the question.

Moreover, we added one question to Task Breach.

(Breach.Recovery: Task 2.5)   "What are the actions/steps taken after the breach has happened? If such (recovery) actions are not stated in the breach report, answer 'Not stated'." In addition to the norm that would prevent the breach by eliminating its cause, the answer to this question would be the consequent of an alternative norm that describes a recovery plan for the aftermath of a breach.

Breach Modifications After analyzing the average quality of the extraction, we have identified five breach reports that received unsatisfactory results. We believe that, if these breach reports are presented in a more concise and structured way, workers may provide more usable responses. We prepared three separate versions (in addition to the original description) for each of the five breach reports. Three authors slimmed down the reports individually with the goal of reducing complexity. We took out sentences that we thought were irrelevant to our survey questions. For example, sentences regarding breach notifications ("The CE provided breach notification to HHS …") or OCR assurances ("OCR obtained assurances that the CE implemented the corrective actions noted above") bear no grave significance in the remedy and prevention of the breaches, and they were contained in the majority of breach reports. We include the original and a modified version of an example breach in the Appendix. The results for these revisions are presented in Section 6.2.

Worker Groups Since we did not find any statistically significant difference between the quality of responses from worker groups, we chose to deploy the final study using qualified workers, i.e., regular workers with qualifications, which was the best performing group on average.

Compensation To encourage and reward better quality responses, we increased the bonus payment to $3.

5.3 Final Study Results

In the final study, we rejected workers that participated in multiple batches or provided clearly incorrect or random answers. In addition, we rejected workers that did not follow our instructions. In this manner, we rejected 11 (23.9%) out of 46 workers. The average quality in the final study was 0.81, 13.9% higher than the average quality of the same worker group (regular workers with qualifications) in the pilot. The details of specific results regarding our research questions are shown below.

RQ$_1$ Figure 5 shows the average quality for each type of breach. The average response quality for each breach type increased over the pilot study. Moreover, there is no significant difference in quality among breach types.

RQ$_2$ With our aforementioned revisions, we expected smaller variance among tasks. Figure 6 shows the average quality of each task in our final study.

We can see that the average quality of responses to Tasks Breach.Condition (Task 2.3) and Regulation.Condition (Task 4.4) was still lower than the response quality for other tasks, but this response quality increased 53% and 39%, respectively. In addition, we observed that the response quality for Task Breach.Action (Task 2.1) notably increased (27%). However, the response quality for Task Regulation.Action (Task 4.3) had a smaller increase (10%), despite having similar instructions.

In the pilot, we observed that many workers responded "Always" or "In every circumstance" to the questions that corresponded to the antecedent of a norm, which was nonspecific and sometimes logically incorrect. In our revised questionnaire, we asked workers to start their answers with "When" or "If." We observed a lot of responses like "When I'm accessing protected health information then I believe yes I always have to take that action." This answer is more acceptable because it includes a valid antecedent ("accessing protected health information") for the action.
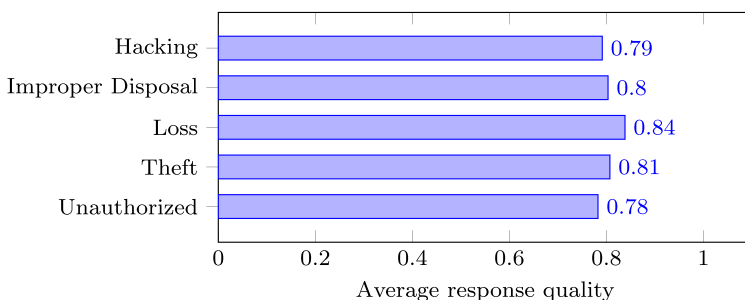

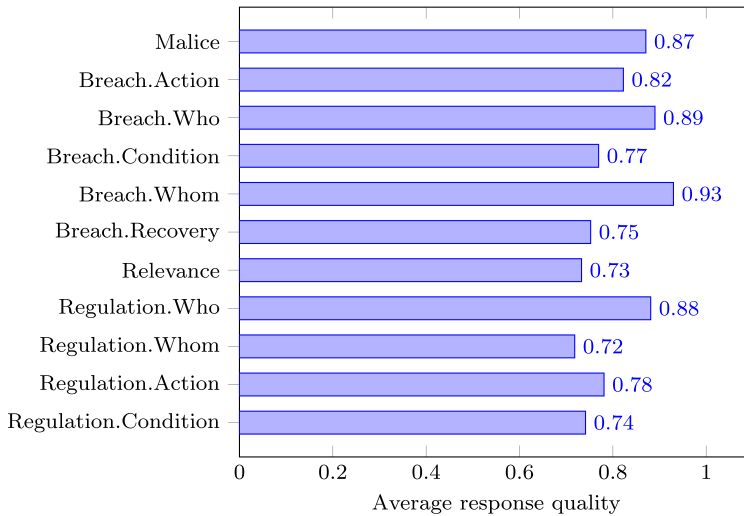**Fig. 5** Performance across breach types in the final study

Fig. 6 Performance across tasks in final study

Worker Performance We also investigated other factors that might have affected the crowd workers' performance, such as perception, experience, and learning curve. However, there was no significant correlation ($p = 0.395 > 0.10$) between the quality of responses and perceived difficulty. Among all workers, 24.5% reported that they had no prior experience in regulatory text. The average quality of their responses was 0.827, which was slightly higher than responses from workers with experience (0.801). We conclude that workers' prior experience does not affect task performance ($p = 0.646 > 0.10$). Also, we sought to understand whether workers performed better as they went further along the survey, e.g., did they perform better for the last two breaches after completing the first three? However, we did not find any significant difference.

Worker Effort In the final study, the average time each worker spent on the tutorial was 4.3 minutes and the median was 3.2 minutes. The mean time each worker spent on one breach was 12.9 minutes and the median was 9.5 minutes. The average time it took each worker to finish the entire study was 68.9 minutes, which was slightly higher than the pilot. This increase is reasonable since we had one more task in the final study.

Evaluation Disagreements Our evaluators had moderate agreement (Kappa = 0.54, $p < 0.001$; 95% confidence interval from 0.503 to 0.578). They disagreed on 16.6% of all evaluated entries, 62% of which came from Task Regulation, especially Task Regulation.Whom (Task 4.2) (26%). The disagreements resulted from the evaluators' different standards or understanding of the responses. They resolved these disagreements through reexaminations and discussions.

5.4 Connecting Breach Reports to HIPAA Clauses

In Task 3, we have asked the crowd workers to rate the relevance of the HIPAA clauses to the breach reports on a scale of 1 to 4 where 1 stands for "not relevant" and 4 stands for "exactly the case in the breach." For each breach report, a worker is asked to rate five

HIPAA clauses, one of which was identified as the most relevant (in previous work Kafalı et al. 2017), and the other four were randomly chosen. The crowd workers rated the most relevant clauses 2.7 on average, and the others 2.3 on average. Details of the ratings can be found in our published dataset.

We have asked the workers to extract requirements from both the breach reports and the most relevant HIPAA clauses. For example, one worker provided "encrypting all flash drives containing electronic protected health information" as the consequent of a norm extracted from a breach report about a loss of a flash drive, whereas the consequent of the norm extracted from the relevant clause was "implement ways to encrypt and decrypt protected health information." The former norm can be considered as a security requirement that is more specific and practical than the latter legal requirement.

For some of the breach reports, the crowd workers rated HIPAA clauses as the most relevant ones that are different from the ones identified in Kafalı et al. (2017). For example, in a Loss type of breach, employees used personal emails for work purposes. The previous study (Kafalı et al. 2017) identified a clause on workstation use (164.310(b)), whereas the workers rated a clause on security management process (164.308(a)(1)(i)).

5.5 From Crowd Responses to Norms

Table 2 shows all responses to Task Breach (Task 2) of the breach report presented in Section 2. On average, each breach report received accepted responses from 4.75 workers. In Table 2, each crowd worker, identified by his/her ID, has given answers to four sub-tasks,

Table 2 Responses to Task Breach of the example breach

ID	Task	Response	Sentences
#312	Action	good	3,5,6,8
	Who	good	2,3,5,8
	Condition	good	3,4,5,6
	Whom	good	2,5,7,8
#323	Action	The portable drive should have been better safe-guarded, including using data encryption	1,7
	Who	The pharmacy resident	1
	Condition	When handling patients' data it should always be encrypted and handled with the utmost concern	1
	Whom	Arnold Palmer Hospital and some its patients who were involved in a study	1,2,3
#333	Action	Data extracted on the given information	6,7
	Who	protected health information	2,3
	Condition	Substitute notice was posted on the CE's website	5,6
	Whom	protected health information	4,7
#344	Action	Provide ample training to residents	7
	Who	Arnold Palmer Hospital	1
	Condition	When training for residents is necessary	7
	Whom	HHS, patients, and babies who were part of a study	2,3,4
#349	Action	Train employees about data loss, data protection	7
	Who	the covered entity	7
	Condition	When PHI is involved … take the action	7
	Whom	patients, the covered entity, the employee involved	1,2

namely, Breach.Action (Task 2.1), Breach.Who (Task 2.2), Breach.Condition (Task 2.3), and Breach.Whom (Task 2.4). Each answer contains a textual response and the indices of sentences that contain the response.

We use the same evaluators to formalize norms from these responses manually as the final results of ÇORBA. Since we have designed the questionnaires in the format of norms, composing norms from the responses is straightforward, which we describe below.

Response #312 contains only random answers, and therefore has been rejected. Response #333 has been rejected because the answers are clearly unreasonable ("protected health information" to "who should have taken the action") and did not follow instructions. Note that Response #323 has been evaluated as acceptable and therefore kept even though the answer to Task Breach.Action did not start with a verb.

From Response #323, we directly obtain the norm c(EMPLOYEE, CE, portable devices contain PHI, safeguard portable devices) or p(EMPLOYEE, CE, portable devices contain PHI, lose portable devices), where CE stands for Covered Entity. This norm resides in Sentence 1. Since the answers are in style of commitment, prohibition type of norms are concluded intuitively for better understanding. From Responses #344 and #349, we directly obtain the norm c(CE, PATIENTS, true, train employees on data loss and data protection). Elements of this norm are extracted from Sentences 2 and 7.

In this manner, the evaluators have manually combined and collected 60 unique norms from the 38 breach reports. Examples given in Section 2 are from this set. We did not perform this process for the regulatory clauses, as we only included the most relevant eight clauses for norm extraction.

6 Discussion

Regulations and breach reports contain useful information that can help developers and security analysts prevent future breaches. Previous work has developed normative models (Barth et al. 2006; Hao et al. 2016; Kafalı et al. 2017; Singh 2013) to formalize such information. ÇORBA is a solution for effective normative information extraction from regulations and breach reports so that further automated reasoning techniques can be applied to help developers and security analysts reach their goals. In this section, we present the key findings of our study, some additional findings regarding breach report revisions, and some insights into automating such revisions.

6.1 Key Findings

As answers to our research question, we present the following key findings regarding the designing of crowdsourcing projects for requirement extraction from regulations and breach reports.

Research Questions By investigating RQ_1, we have not found significant correlations between the extraction performance and report lengths, complexities, perceived difficulties, and type of breach. By investigating RQ_2, we have found that crowd workers give lower quality answers to tasks the required actions, and the conditions under which the actions are required. The format of the survey question can have a considerable influence on their

responses. With instructions that are more detailed and tailored, the responses can be dramatically improved. We have shown that personalized questions (e.g., "Consider you are the responsible person for carrying out . . . "), reminders (e.g., "Please keep in mind . . . "), and specific instructions (e.g., "Start your answer with a verb") can improve response quality from workers.

Security Requirements The resulting norms extracted from the breach reports can be considered as security requirements for socio-technical systems (Dalpiaz et al. 2016) in the healthcare domain. The norms include software requirements (e.g., "conduct file transfers through a secured network" and "store PHI only in encrypted programs") and organizational policies (e.g., "validate contents of letters before mailing" and "digitize and destroy hard copy records"). In addition, norms extracted from a breach report correspond to breach prevention, the violation of which may lead to similar incidents. This information could be of great help to security analysts.

Crowdsourcing For a non-typical crowdsourcing study like ours, we need to set up appropriate tasks and rewards. To improve the quality of the worker responses, we provided detailed instructions in the final study, and rejected the workers that did not follow instructions. The reward promised to the workers was much higher than average projects on mTurk. We provided fitting rewards and bonuses based on worker effort (time spent) and behavior, as an incentive, to ensure the quality of their responses.

Deliverable We have created a curated dataset with evaluated worker responses. This dataset can be used for future research to train automated tools on mining normative elements. Our dataset contains 2,850 worker answers from the pilot, and 3,360 answers from the final study. All responses have been evaluated by two evaluators. Thus, our dataset includes a total of 6,210 evaluated answers. From these responses, we have concluded 60 refined norms. Extracted norms are also connected to the selected HIPAA clauses with relevance gradings from workers. We have made this dataset public alongside all our study material here: https://goo.gl/xda2nQ.

6.2 Results from Breach Report Revisions

Through investigation of all breaches contained in the HHS breach reports as well as evaluation of worker responses for a subset of the breaches, we have identified that breach reports self-reported by end users often contain irrelevant information for security requirements engineering. Table 3 summarizes the average response quality across the breach versions that we have created for the final study as well as the original descriptions used in both studies. Workers in the final study performed remarkably better for those breaches than workers in the pilot. Moreover, one of our breach versions (v2) yielded higher quality responses (outperforming others on average and for three of the breaches). Our revisions focused on the shortening of the breach reports by keeping only the sentences that contained key information.

Our results have shown that most revised versions have outperformed the original versions, and this kind of revisions can facilitate more effective information extraction. Whereas this finding is promising, further guidelines and templates for creating structured

Table 3 Performance across breach versions

Breach	Original (Pilot)	Original (Final)	v1	v2	v3
#11	0.42	0.73	0.8	0.96	0.55
#27	0.67	0.74	0.64	0.83	0.78
#182	0.55	0.67	0.81	0.86	0.81
#495	0.6	0.72	0.86	0.79	0.83
#657	0.59	0.83	0.72	0.89	0.93
Average	0.57	0.74	0.77	0.87	0.78

Boxes mark best performance in a row

natural language documents would be helpful for producing good quality requirements (Arora et al. 2015). Continuing this work could be an interesting direction toward new standards such as those developed by National Institute of Standards and Technology (NIST) for vulnerabilities (CVSS) and misuses (CMSS), which facilitate data extraction and intra- and inter-organizational analysis.

6.2.1 Automated Revisions

Automated methods can be leveraged to facilitate the aforementioned revisions. Our curated dataset contains extracted norms from the breach reports as well as the sentences from which the norm elements are extracted. Using this information as a training set, we can use a probabilistic binary classifier to identify the sentences that contain useful information toward norm extraction.

We have experimented this process using Paragraph Vectors (Le and Mikolov 2014), also known as Doc2Vec, for document embedding and a probabilistic classifier, Logistic Model Trees (Landwehr et al. 2005; Sumner et al. 2005), for classification. Textbox 3 shows the automatic ranking of sentences in Textbox 2 based on their probabilities of containing useful information. The numbers leading the sentences mark the original order of the sentences.

Using 0.50 as a threshold, the elements of the refined norms are from Sentences 1, 2 and 7, which have been correctly identified. Sentence 6 is also likely to contain a norm (implementing a data loss prevention system). Sentence 3 contains information regarding the affected patients (babies). Sentences 4, 5 and 8 are deemed to be less important. In fact, they were removed in at least two revised versions. Using 10-fold cross validation, the accuracy of this classification is 76.6% (using 0.50 as a threshold).

This process ranks each sentence in a report by its relevance in norm extraction. We can use this result to revise new breach reports by keeping only relevant sentences. To improve the recall of retrieving relevant sentences, we can use a lower threshold, e.g., 0.25, which will give us a recall of 91.1% with moderate precision (56.7%). Alternatively, we can highlight the sentences in the reports with different colors corresponding to their importance, which could facilitate crowd workers in the extraction.

Textbox 3

P=0.95 (1) An unencrypted portable data drive was lost by a pharmacy resident of the Arnold Palmer Hospital, a part of the covered entity (CE).

P=0.89 (6) The CE updated its policies and procedures for its data loss prevention system and added controls.

P=0.76 (3) The missing drive also stored personal items, a research study proposal, and two spreadsheets containing limited information on 586 babies who were part of a study.

P=0.55 (7) The CE retrained the resident involved in the loss of data and provided additional information to all employees and medical staff members regarding the use of portable data devices through education and published articles.

P=0.50 (2) The drive contained the protected health information (PHI) of 586 individuals, including names, birth weights, gestational age, admission and discharge dates, medical record numbers, and some transfer dates.

P=0.36 (5) Substitute notice was posted on the CE's website.

P=0.18 (4) The CE provided breach notification to HHS, the media, and to the parents of the affected individuals because they were all minors.

P=0.06 (8) OCR obtained assurances that the CE implemented the corrective actions listed above.

---

7 Threats to Validity

In this section, we discuss the threats to validity of our methodology.

Generalizability and Scalability Overall, we employed 80 workers in our study. Workers spent longer on our HIT that a typical one and had to work on several questionnaires. We collected 6,210 individual responses from the workers. Analysis of such data requires considerable evaluation effort. Still, this sample may not be large enough to draw general conclusions about information extraction tasks. Moreover, we crowdsourced the analysis of 38 breach reports, which is a small fraction of approximately 1,000 reports. We attempted to cover different types of breaches with our selection, but other reports may include unique information that needs individual examination.

To scale the extraction up to larger datasets, we need to deploy more batches of the same questionnaires, which requires additional financial and time costs. In our study, each batch took approximately three dollars per report and five days per batch on average. However, we do not think this is a major issue with current number of available breach reports (about two thousand on HHS website as of 2018). It is practical to increase our dataset using the current methodology, and integrate the automation process when the dataset is reasonably large.

Also, we need to conduct extensive studies to validate the drafting guideline. In addition, our study has been limited to HIPAA and HHS breach reports. We need to evaluate our methodology by applying it to other security related textual artifacts.

Question Format We have shown that the way in which a question is framed can affect the quality of the responses. In our questionnaire, especially in the pilot, we focused on the norms the violation of which led to the breaches. Specifically, we asked for the necessary actions to prevent the breaches. However, breach reports additionally include useful information regarding actions that should be taken after a breach, e.g., retraining of staff and notifications to HHS and media. In the final study, we added Task Breach.Recovery (Task 2.5) to mitigate this problem. One breach report may contain multiple norms that are valid and important, which we should take into account when designing such a questionnaire. In addition, our questions are predominantly asked in a style of commitment norms since the HHS breach reports we adopted mostly concern undesired disclosure of information. This limitation leaves room for improvement since other settings may involve situations where a denial of access is a violation of an authorization, which we retain in this work.

Legal Expertise None of the authors have expertise to investigate HIPAA from a legal perspective. We have not consulted a legal expert to validate the selection of the HIPAA clauses that we identify as the most relevant to the associated breach reports. We have conducted our studies on a limited collection of HIPAA clauses on security and privacy, each of which was shown to the crowd workers verbatim. We have found that the clauses are fairly easy to understand and do not present a large amount of ambiguities. To further evaluate the effectiveness of crowdsourcing on elicitation of regulatory requirements, we must take ambiguities in legal texts (Massey et al. 2014) into consideration.

8 Conclusions and Future Work

We have presented ÇORBA, a methodology for extracting and connecting useful information, regarding the obtaining of security and privacy requirements, from regulations and breach reports using crowdsourcing, as well as its evaluation on automated methods. Using our methodology, we have deployed a crowdsourcing study on Amazon Mechanical Turk (mTurk). Capturing such information would enable the specification of formal normative models upon which automated reasoning can be performed to help software developers and security analysts.

We have showed that the crowdsourcing methodology, taking into account proper question formats, instructions, and text revisions, can yield extraction results with high quality. By leading the crowd workers to provide proper answers with desired formats, we can significantly improve the quality of the responses. We have created a curated dataset that contains 6,210 evaluated worker responses performed by two evaluators. This dataset will be helpful for future research to train natural language processing methods as well as to design similar crowdsourcing tasks.

Future work includes additional methods to improve worker responses, such as contest type tasks (Dwarakanath et al. 2016) and alternative worker compensations. We will explore the use of automated extraction tools to complementing the crowdsourcing methodology. For example, we can automatically highlight parts of text that have a high probability of containing a requirement, thus helping the crowd worker identify valuable knowledge. To do so, we can use heuristics as well as sequence labeling techniques, such as part-of-speech

tagging, to find each element from a sentence that contains a norm. Moreover, to validate the applicability of ÇORBA in other domains, we plan to investigate the Verizon Data Breach Investigations Reports (Verizon 2016), the DataLoss Database (DataLossDB 2015), and the Principedia privacy incidents database (Staddon 2016).

Appendix

A.1 Breach Modifications

The original breach report for Breach #11 from Table 3 is shown below.

"On or around June 15, 2012, an employee of the covered entity (CE), Advanced Data Processing, Inc. (ADP), dba Intermedix, who had access to patients' protected health information (PHI) as part of her job, inappropriately accessed the PHI of approximately 10,000 individuals and sold the information to third parties. An addendum to the initial breach report, submitted on April 3, 2015, expanded the breach to an additional 2,360 individuals. The PHI involved in the breach included patient names, social security numbers, addresses, dates of birth, claims, and other financial information. The CE provided breach notification to HHS, affected individuals, and the media and posted substitute notice. Following the breach, the CE engaged a third party to review its network environment and make recommendations for security enhancements. It implemented data loss prevention technology to identify electronic PHI and block transmittal of sensitive information and a log management and analysis solution to automate collection, analysis, archival and recovery of log data. The CE implemented policies and procedures for disposal and reuse of mobile devices, as well as for the secure transport of sensitive information to, from, and between data centers. The CE also created an information security team and appointed a committee to address compliance. Additionally, the CE improved its employee training program and launched a vendor management program to ensure the safeguarding of ePHI by its business associates. OCR obtained assurances that the CE implemented the correction actions listed above. The CE also initiated upgrades to its data center security and workstation antivirus technology."

The modified breach report v2 for Breach #11 from Table 3 is shown below.

"On or around June 15, 2012, an employee of the covered entity (CE), Advanced Data Processing, Inc. (ADP), dba Intermedix, who had access to patients' protected health information (PHI) as part of her job, inappropriately accessed the PHI of approximately 10,000 individuals and sold the information to third parties. The PHI involved in the breach included patient names, social security numbers, addresses, dates of birth, claims, and other financial information. Following the breach, the CE engaged a third party to review its network environment and make recommendations for security enhancements. It implemented data loss prevention technology to identify electronic PHI and block transmittal of sensitive information and a log management and analysis solution to automate collection, analysis, archival and recovery of log data. The CE implemented policies and procedures for disposal and reuse of mobile devices, as well as for the secure transport of sensitive information to, from, and between data centers. The CE also created an information security team and appointed a committee to address compliance. Additionally, the CE improved its employee training program and launched a vendor management program to ensure the safeguarding of ePHI by its business associates."

A.2 Survey Tutorial

Figure 7 shows what the workers see in the study tutorial as a sample "correct" answer for Task Malice (Task 1). Figure 8 shows what the workers see in the study tutorial as sample "correct" answers for some of the questions under Task Breach (Task 2).

A.3 Worker Responses

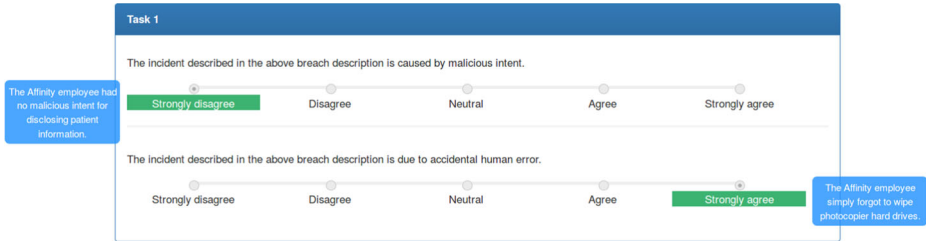Figure 9 shows answers from a sample worker for Task Regulation (Task 4).
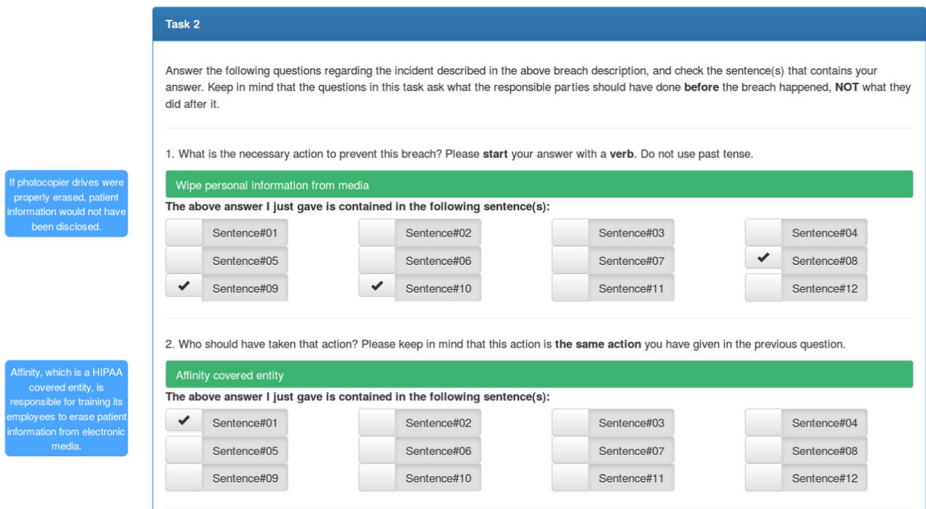

Fig. 7 Survey tutorial: Task Malice (Task 1)

Fig. 8 Survey tutorial: Part of Task Breach (Task 2)

Fig. 9 Worker Response: Task Regulation (Task 4)

References

Allen IE, Seaman CA (2007) Likert scales and data analyses. Qual Prog 40(7):64

Arora C, Sabetzadeh M, Briand L, Zimmer F (2015) Automated checking of conformance to requirements templates using natural language processing. IEEE Trans Softw Eng 41(10):944–968

Barth A, Datta A, Mitchell JC, Nissenbaum H (2006) Privacy and contextual integrity: framework and applications. In: Proceedings of the IEEE symposium on security and privacy (SP). IEEE Computer Society, Washington, DC, pp 184–198

Bhatia J, Breaux TD, Schaub F (2016) Mining privacy goals from privacy policies using hybridized task recomposition. ACM Transa Softw Eng Methodol (TOSEM) 25(3):1–24

Breaux TD, Antón AI (2008) Analyzing regulatory rules for privacy and security requirements. IEEE Trans Softw Eng 34(1):5–20

Breaux TD, Schaub F (2014) Scaling requirements extraction to the crowd: experiments with privacy policies. In: Proceedings of the 22nd international requirements engineering conference (RE), pp 163–172

Dalpiaz F, Paja E, Giorgini P (2016) Security requirements engineering: designing secure socio-technical systems. The MIT Press

Dam HK, Savarimuthu BTR, Avery D, Ghose A (2015) Mining software repositories for social norms. In: Proceedings of the 37th international conference on software engineering (ICSE). IEEE Press, pp 627–630

DataLossDB (2015) 2015 reported data breaches surpasses all previous years. https://blog.datalossdb.org/2016/02/11/2015-reported-data-breaches-surpasses-all-previous-years/

Dean D, Gaurino S, Eusebi L, Keplinger A, Pavlik T, Watro R, Cammarata A, Murray J, McLaughlin K, Cheng J et al (2015) Lessons learned in game development for crowdsourced software formal verification. In: Proceedings of USENIX summit on gaming, games, and gamification in security education (3GSE 15). USENIX Association, Washington, D.C

Downs JS, Holbrook MB, Sheng S, Cranor LF (2010) Are your participants gaming the system?: screening mechanical turk workers. In: Proceedings of the SIGCHI conference on human factors in computing systems CHI '10. ACM, New York, pp 2399–2402

Dwarakanath A, Shrikanth NC, Abhinav K, Kass A (2016) Trustworthiness in enterprise crowdsourcing: a taxonomy & evidence from data. In: Proceedings of the 38th international conference on software engineering companion. ACM, pp 41–50

Gao X, Singh MP (2014) Extracting normative relationships from business contracts. In: Proceedings of the 13th international conference on autonomous agents and multiagent systems (AAMAS). IFAAMAS, Paris, pp 101–108

Getman AP, Karasiuk VV (2014) A crowdsourcing approach to building a legal ontology from text. Artif Intell Law 22(3):313–335

Ghanavati S, Rifaut A, Dubois E, Amyot D (2014) Goal-oriented compliance with multiple regulations. In: Proceedings of IEEE 22nd international requirements engineering conference (RE), pp 73–82

Gürses S, Rizk R, Günther O (2008) Privacy design in online social networks: learning from privacy breaches and community feedback. In: Proceedings of international conference on information systems (ICIS), p 90

Hao J, Kang E, Sun J, Jackson D (2016) Designing minimal effective normative systems with the help of lightweight formal methods. In: Proceedings of the 24th ACM SIGSOFT international symposium on the foundations of software engineering (FSE), pp 50–60

Hashmi M (2015) A methodology for extracting legal norms from regulatory documents. In: Proceedings of IEEE 19th international enterprise distributed object computing workshop, pp 41–50

HHS (2003) Summary of the HIPAA privacy rule. United States Department of Health and Human Services (HHS). http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/

HHS Breach Portal (2016) Notice to the Secretary of HHS breach of unsecured protected health information affecting 500 or more individuals. United States Department of Health and Human Services (HHS). https://ocrportal.hhs.gov/ocr/breach/

Kafalı Ö, Ajmeri N, Singh MP (2016a) Revani: revising and verifying normative specifications for privacy. IEEE Intell Syst 31(5):8–15

Kafalı Ö, Singh MP, Williams L (2016b) Nane: identifying misuse cases using temporal norm enactments. In: Proceedings of the 24th IEEE international requirements engineering conference (RE). IEEE Computer Society, Beijing, pp 136–145

Kafalı Ö, Jones J, Petruso M, Williams L, Singh MP (2017) How good is a security policy against real breaches? a HIPAA case study. In: Proceedings of the 39th international conference on software engineering (ICSE). IEEE Computer Society, Buenos Aires, pp 530–540

Kashyap A, Han L, Yus R, Sleeman J, Satyapanich T, Gandhi S, Finin T (2016) Robust semantic text similarity using LSA, machine learning, and linguistic resources. Lang Resour Eval 50(1):125–161

Landwehr N, Hall M, Frank E (2005) Logistic model trees. Mach Learn 59(1–2):161–205

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: Proceedings of the 31st International conference on international conference on machine learning - vol 32, ICML'14, pp 1188–1196

Liu Y, Sarabi A, Zhang J, Naghizadeh P, Karir M, Bailey M, Liu M (2015) Cloudy with a chance of breach: forecasting cyber security incidents. In: Proceedings of the 24th USENIX conference on security symposium, pp 1009–1024

MacLean DL, Heer J (2013) Identifying medical terms in patient-authored text: a crowdsourcing-based approach. J Am Med Inform Assoc 20(6):1120–1127

Massey AK, Rutledge RL, Antón AI, Swire PP (2014) Identifying and classifying ambiguity for regulatory requirements. In: 2014 IEEE 22nd international requirements engineering conference (RE), pp 83–92

Matulevičius R, Mayer N, Heymans P (2008) Alignment of misuse cases with security risk management. In: Proceedings of the 3rd international conference on availability, reliability and security (ARES), pp 1397–1404

Maxwell JC, Anton AI (2009) Developing production rule models to aid in acquiring requirements from legal texts. In: 2009 17th IEEE International requirements engineering conference, pp 101–110

Murukannaiah PK, Ajmeri N, Singh MP (2016) Acquiring creative requirements from the crowd: understanding the influences of individual personality and creative potential in crowd RE. In: Proceedings of the 24th IEEE international requirements engineering conference (RE). IEEE Computer Society, Beijing, pp 176–185

Murukannaiah PK, Dabral C, Sheshadri K, Sharma E, Staddon J (2017) Learning a privacy incidents database. In: Proceedings of the hot topics in science of security: symposium and bootcamp, HoTSoS. ACM, New York, pp 35–44

Patwardhan M, Sainani A, Sharma R, Karande S, Ghaisas S (2018) Towards automating disambiguation of regulations: using the wisdom of crowds. In: Proceedings of the 33rd ACM/IEEE international conference on automated software engineering, pp 850–855

Reidenberg JR, Breaux T, Carnor LF, French B (2015) Disagreeable privacy policies: mismatches between meaning and users' understanding. Berkeley Technol Law J 30(1):39

Riaz M, King J, Slankas J, Williams L (2014) Hidden in plain sight: automatically identifying security requirements from natural language artifacts. In: Proceedings of the 22nd IEEE international requirements engineering conference (RE), pp 183–192

Riaz M, Stallings J, Singh MP, Slankas J, Williams L (2016) DIGS: a framework for discovering goals for security requirements engineering. In: Proceedings of the 10th ACM/IEEE international symposium on empirical software engineering and measurement (ESEM). ACM, pp 35:1–35:10

Savarimuthu BTR, Dam HK (2014) Towards mining norms in open source software repositories. In: Agents and data mining interaction, lecture notes in computer science, vol 8316. Springer, Berlin, pp 26–39. https://doi.org/10.1007/978-3-642-55192-5_3

Siena A, Jureta I, Ingolfo S, Susi A, Perini A, Mylopoulos J (2012) Capturing variability of law with Nómos 2. In: Atzeni P, Cheung D, Ram S (eds) Conceptual modeling. Springer, Berlin, pp 383–396

Sindre G, Opdahl AL (2005) Eliciting security requirements with misuse cases. Requir Eng 10(1):34–44

Singh MP (2013) Norms as a basis for governing sociotechnical systems. ACM Trans Intell Syst Technol (TIST) 5(1):21,1–21,23

Slankas J, Williams L (2013) Access control policy extraction from unconstrained natural language text. In: Proceedings of the international conference on social computing (SocialCom), pp 435–440

Sleimi A, Sannier N, Sabetzadeh M, Briand L, Dann J (2018) Automated extraction of semantic legal metadata using natural language processing. In: Proceedings of IEEE international requirements engineering conference (RE), pp 124–135

Staddon J (2016) Privacy incidents database: the data mining challenges and opportunities. Cyber Security Practitioner

Sumner M, Frank E, Hall M (2005) Speeding up logistic model tree induction. In: Proceedings of the 9th European conference on principles and practice of knowledge discovery in databases. Springer, Berlin, pp 675–683

Verizon (2016) Data breach investigations reports. http://www.verizonenterprise.com/verizon-insights-lab/dbir/

Von Wright GH (1999) Deontic logic: a personal view. Ratio Juris 12(1):26–38

Wilson S, Schaub F, Ramanath R, Sadeh N, Liu F, Smith NA, Liu F (2016) Crowdsourcing annotations for websites' privacy policies: can it really work? In: Proceedings of the 25th international conference on world wide web. International World Wide Web Conferences Steering Committee, pp 133–143

Zeni N, Kiyavitskaya N, Mich L, Cordy JR, Mylopoulos J (2015) GaiusT: supporting the extraction of rights and obligations for regulatory compliance. Requir Eng 20(1):1–22

Zeni N, Mich L, Mylopoulos J (2017) Annotating legal documents with GaiusT 2.0. Int J Metadata Semant Ontol 12:47

Zeni N, Seid EA, Engiel P, Mylopoulos J (2018) NómosT: building large models of law with a tool-supported process. Data Knowl Eng 117:407–418

Hui Guo is a PhD student in Computer Science at NC State University. His research interests include multi-agent systems, NLP, text mining, and crowdsourcing. Guo has an MS in Computer Science from East Carolina University, and a BS in Automation Engineering from Tsinghua University.

Dr. Özgür Kafalı is a Lecturer and the Ethics Officer at the School of Computing at University of Kent, UK. His research interests lie within the intersection of multiagent systems, cybersecurity, and requirements engineering. Kafalı received a PhD in Computer Engineering from Bogazici University, Turkey.

Anne-Liz Jeukeng is a Software Engineer on a cybersecurity team at the Aerospace and Defense company Northrop Grumman. Jeukeng is also obtaining her MS Degree in Computer Science from the University of Massachusetts Dartmouth. Jeukeng has a BS in Computer Science from University of Florida. She is fascinated by cybersecurity especially user safety.

Dr. Laurie Williams is a Distinguished Professor in the Computer Science department North Carolina State University (NCSU) and an IEEE Fellow. Laurie has been the co-director of the NSA-sponsored NCSU Science of Lablet research center since 2011. Laurie's research focuses on software security; agile software development practices and processes, particularly continuous deployment; and software reliability.

Dr. Munindar P. Singh is a Professor in Computer Science and a co-director of the Science of Security Lablet at NC State University. His research interests include the engineering and governance of sociotechnical systems. Singh is an IEEE Fellow, a AAAI fellow, and a former Editor-in-Chief of IEEE Internet Computing and the ACM Transactions on Internet Technology.

Affiliations

Hui Guo[1] (iD) · Özgür Kafalı[2] · Anne-Liz Jeukeng[3] · Laurie Williams[1] ·
Munindar P. Singh[1]

Özgür Kafalı
R.O.Kafali@kent.ac.uk

Anne-Liz Jeukeng
lizjeukeng@gmail.com

Laurie Williams
lawilli3@ncsu.edu

Munindar P. Singh
mpsingh@ncsu.edu

[1]   Department of Computer Science, North Carolina State University, Raleigh, NC, USA

[2]   School of Computing, University of Kent, Canterbury, UK

[3]   Department of Computer & Information Science & Engineering, University of Florida,
      Gainesville, FL, USA