Poster: Toward Extraction of Security Requirements from Text

Hui Guo
North Carolina State University
Raleigh, North Carolina, USA
hguo5@ncsu.edu

Özgür Kafalı
University of Kent
Canterbury, Kent, UK
R.O.Kafali@kent.ac.uk

Anne-Liz Jeukeng
University of Florida
Gainesville, Florida, USA
anneliz1@ufl.edu

Laurie Williams
North Carolina State University
Raleigh, North Carolina, USA
lawilli3@ncsu.edu

Munindar P. Singh
North Carolina State University
Raleigh, North Carolina, USA
mpsingh@ncsu.edu

ABSTRACT

We propose and evaluate an information extraction and analysis framework that combines human intelligent (crowdsourcing) with automated methods to produce improved security and privacy requirements incorporating knowledge from post-deployment artifacts such as breach reports.

CCS CONCEPTS

• **Information systems** → **Crowdsourcing**; • **Social and professional topics** → *Socio-technical systems*; • **Computing methodologies** → Natural language processing;

KEYWORDS

Regulatory norms, sociotechnical systems, crowdsourcing, natural language processing, HIPAA

ACM Reference Format:
Hui Guo, Özgür Kafalı, Anne-Liz Jeukeng, Laurie Williams, and Munindar P. Singh. 2018. Poster: Toward Extraction of Security Requirements from Text. In *HoTSoS '18: Hot Topics in the Science of Security: Symposium and Bootcamp, April 10–11, 2018, Raleigh, NC, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3190619.3191685

1 GOAL

In domains such as healthcare, regulations, e.g., HIPAA[1], specify how software and its users should behave in security-critical situations. However, regulations are often arcane and unclear as to actionable requirements. Breach reports not only describe cases where deployed systems fail or are maliciously or accidentally misused, but also suggest actions to prevent, detect, and recover from future breaches. Such knowledge could be highly valuable in refining system requirements, in essence, by supplementing the applicable regulations. *Our research goal is to produce improved security and privacy requirements that (1) accommodate social, not just technical, considerations and (2) incorporate knowledge from post-deployment artifacts such as breach reports.*

HoTSoS '18, April 10–11, 2018, Raleigh, NC, USA

ACM ISBN 978-1-4503-6455-3/18/04.
https://doi.org/10.1145/3190619.3191685

2 METHOD

Accordingly, we propose a framework that combines crowdsourcing and automated methods to extract a structured normative formalization from textual artifacts. The framework incorporates data collection, evaluation, and automated methods.

For data collection, we extract a structured normative formalization from textual artifacts using crowdsourcing where workers respond to survey style questions. We address research questions regarding the factors that affect the quality of worker responses, including worker types, worker experience, and question format. We carefully construct the questions and perform this step in two phases. Based on the results from Phase I, we adjusted the survey questions, instructions, and project settings for Phase II. We validated our crowdsourcing methodology for HHS breach reports [2] via Amazon Mechanical Turk (mTurk).

For evaluation, we aggregate the evaluations of multiple human evaluators to determine the quality of the responses and to refine the extracted norms. Our results show that responses from Phase II were of significantly higher quality than those from Phase I. We have created a curated dataset with evaluated worker responses, from which we manually extracted well-structured norms. This dataset can be used for future research to train automated tools on mining normative elements.

For automation, we leverage results from the previous two steps in automatic extraction and classification of norms to increase the scale and efficiency of the extraction process. We demonstrate that classification techniques offer high performance in identifying the existence and types of norms in a given sentence. For a sentence that is deemed to contain a norm, semantic similarity can be incorporated to retrieve the most similar norms in the dataset to improve the efficiency of the extraction process.

3 CONCLUSIONS AND FUTURE WORK

We present and evaluate a framework for extracting knowledge from regulations and breach reports using crowdsourcing and automated methods. Future work includes different compensation strategies, multilevel reviews, and fully automated extraction tools.

REFERENCES

[1] HHS. 2003. Summary of the HIPAA privacy rule. (2003). HHS. http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/.
[2] HHS Breach Portal. 2016. Notice to the Secretary of HHS Breach of Unsecured Protected Health Information Affecting 500 or More Individuals. (2016). HHS. https://ocrportal.hhs.gov/ocr/breach/.