ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Master's Thesis Proposal

Query Tool for Analyzing Privacy Changes of GitHub-Hosted OSS Projects

Supervisors: Karel Kubicek, Aileen Nielsen
Professor: Prof. David Basin
Issue Date: November 26, 2021

1 Motivation

How does law become source code? How do programmer communities decide when and how to comply with law, and do they do so correctly? What are the social and technical dynamics that drive or stall legal compliance? These questions have long interested legal scholars, but until recently there have not been opportunities to understand the social and technical dynamics of how law directly affects how and when computer code is written.

There is now an opportunity to address these fundamental questions through an empirical study of two emerging data protection regimes, the General Data Protection Regulation promulgated by the European Union (GDPR) and the California Consumer Privacy Act (CCPA). Much of the data and associated source code for data processing that is governed by GDPR or CCPA is proprietary and therefore cannot be examined directly. However, publicly accessible code deployed for a variety of uses, including commercial purposes, is available in open source repositories on GitHub. An empirical study of trends reflected in GitHub could be highly informative for understanding what prompts programmers to make efforts to comply with data protection laws and how such laws are interpreted when they are implemented in source code.

This study will employ an empirical methodology to identify and analyze open source repositories that address GDPR and CCPA. With Github's API, it is possible to track both the timing and content of code modifications that address GDPR and CCPA. It is also possible to analyze the technical (code dependencies) and social (overlapping contributor status) routes to compliance observed among diverse projects. With these and other empirical measurements developed from the querying tool, it will be possible to address questions of interest to legal scholars about the interplay between law and technology generally, and between legislators and developers specifically.

2 Objective

In this thesis, your task is to implement a tool that allows scanning GitHub for privacy-related interactions. You will collect the commits, pull requests, and issues and reconstruct the communication graph from their references. Your tool will then allow querying this graph for certain patterns of interaction.

You will demonstrate the practicality of the project on reproducing previous manual work. You can also propose new queries that help answers the questions from the motivation. These insights will in turn help to address both legal and economics questions about innovation and legal compliance in technical communities.

3 Tasks

1. Read the existing work [1] on reconstructing GitHub interaction and identify, if they are suitable for our goals.
2. Develop a scraper that collects the issues, commits, and pull requests from GitHub.
3. Reconstruct the interaction graph from 2. using references, code dependencies, and the assignment to a repository and user.
4. Develop querying capability for the interaction graph. Consider the following example queries.
   a) Find all repositories that mentioned first "GDPR" and then "CCPA."
   b) Find issues of project A mentioning GDPR and referencing an issue of project B also mentioning GDPR, where project A includes project B.
   c) Identify large scale events in which a widely shared dependency resulted in identical GDPR or CCPA compliance events across a wide diversity of downstream projects
   d) Identify examples of a single contributor leading compliance efforts in diverse and unrelated projects
5. Evaluate the tool by implementing queries provided by supervisors and propose new queries.
6. Exceptional students can also include other sources of information, such as reported GDPR fines, discussions of privacy enthusiasts, or other communication channels of developers.

4 Deliverables

- At the end of the second week, a time schedule of the project must be shared with the supervisor.
- At the end of the project a presentation of 20 minutes must be given during an InfSec group seminar. It should give an overview as well as the most important details of the work.
- Software, configuration scripts, and any supplementary material must be delivered to the supervisors.
- A final report consisting of an introduction, a discussion on the related work, overview of the procedure and implementation, data from experiments, and the evaluation of common violations. Three copies of this report must be delivered to the supervisor.

References

[1] Aron Fiechter, Roberto Minelli, Csaba Nagy, and Michele Lanza. Visualizing github issues. In *2021 Working Conference on Software Visualization (VISSOFT)*, pages 155–159. IEEE, 2021.

[2] Jérôme Hergueux and Samuel Kessler. Follow the leader: Technical and inspirational leadership in open source software. *Center for Law & Economics Working Paper Series*, 2021(01), 2021.