

Understanding issues related to personal data and data protection in open source projects on GitHub

Anne Hennig
anne.hennig@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Lukas Schulte
lukas.schulte@uni-passau.de
University of Passau
Passau, Germany

Steffen Herbold
steffen.herbold@uni-passau.de
University of Passau
Passau, Germany

Oksana Kulyk
okku@itu.dk
IT University of Copenhagen
Copenhagen, Denmark

Peter Mayer
mayer@imada.sdu.dk
University of Southern Denmark
Odense, Denmark
Karlsruhe Institute of Technology
Karlsruhe, Germany

ABSTRACT

Context: Data protection regulations such as the GDPR and the CCPA affect how software may handle the personal data of its users and how consent for handling of such data may be given. Prior literature focused on how this works in operation, but lacks a perspective of the impact on the software development process.

Objective: Within our work, we will address this gap and explore how software development itself is impacted. We want to understand which data protection-related issues are reported, who reports them, and how developers react to such issues.

Method: We will conduct an exploratory study based on issues that are reported with respect to data protection in open source software on GitHub. We will determine the roles of the actors involved, the status of such issues, and we use inductive coding to understand the data protection issues. We qualitatively analyze the issues as part of the inductive coding and further explore the reasoning for resolutions. We quantitatively analyze the relation between the roles, resolutions, and data protection issues to understand correlations.

CCS CONCEPTS

• **Computing methodologies** → Reasoning about belief and knowledge; Causal reasoning and diagnostics; • **Software and its engineering** → Software creation and management.

ACM Reference Format:

Anne Hennig, Lukas Schulte, Steffen Herbold, Oksana Kulyk, and Peter Mayer. 2023. Understanding issues related to personal data and data protection in open source projects on GitHub. In *Proceedings of International Conference on Mining Software Repositories (MSR '23)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MSR '23, Mon 15 - Tue 26, Melbourne, Australia

© 2023 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Multiple governments across the globe have enacted stricter data protection laws in recent years. The most notable examples are the EU's "Privacy and Electronic Communications Directive 2002/58/EC" (ePrivacy Directive or ePD), which was passed in 2002 and amended in 2009 [2], the General Data Protection Regulation (GDPR) by the European Union (EU), which came into effect in May 2018 [3], and the California Consumer Privacy Act (CCPA) of 2018 [1]. At the core of these regulations are rules that strengthen data protection and force businesses to require valid consent by users for collecting and processing their personal data. Some of the most visible examples are cookie banners. Since these rules are fairly new, their interpretation and implementation is still evolving. For example, when the ePrivacy Directive was initially enacted, many websites only informed users that they collect data using cookies without allowing options to dissent.

With the European Court of Justice's (ECJ) "Planet49" decision (judgment of 1.10.2019 - C-673/17) it became clear that consent for data collection is not freely given, affirmative, informed, and, thus, valid if an opt-out design is used, i.e., via pre-ticked boxes on an online subscription form or – in the case of cookie disclaimers – pre-selected "Agree" options, which users have to actively deselect to refuse consent [5]. Further regulations that had an effect on the design of cookie banners were given in the "Orange România" decision of the ECJ (judgment of 11.11.2020 - C-61/19). It was stated that the free decision of users is disproportionately constrained if the refusal of consent represents a greater effort than the granting of consent. This meant that, e.g., cookie banners where dissent options are hidden in a text or less visible than consent options, can be considered non-compliant with the GDPR [4]. This is just one example how developers, in particular those who offer or maintain a website, needed to constantly keep up with changes due to new regulations over the years.

There is already literature on how software development fails to comply to regulations (e.g., [7, 11, 15, 20, 30, 36, 48]) and how developers discuss privacy related topics (e.g., [19, 34, 46]). These studies have in common that they consider the impact of data protection regulations on software *in operation* or consider data protection discussions *in general*. What we are still missing is a

perspective of how data protection regulations affect the software *development* itself. Software products, for example cookie banners, as described before, may need to be constantly updated to provide new features that enable compliance with data protection laws.

In addition to related work, we want to combine the collection of software that fails to comply to data protection and privacy regulations, *and* also understand how privacy related topics are discussed, not only for already implemented software, but especially also during the software development process. We want to understand who reports issues, what kind of issues are reported and how these reported issues are discussed. Following the course of the discussions, we also want to understand, if reported issues are implemented or - if not - why they are not implemented. To the best of our knowledge, this has not been studied before. Therefore, within this exploratory study, we want to shed light on the impact of data protection regulations like the GDPR on software development on GitHub. Our focus is on the reporting of issues related to personal data and data protection, e.g., change requests, questions, or problems with respect to data protection. In summary, we want to understand *how often* such topics are reported, *what* is reported, *who* reports data protection issues, and what *the reaction* of developers to such issues is. The contributions of our planned research project are the following:

- Insights into the kind of data protection concerns that are reported for open source software and the actors that report them.
- Insights into the actions that developers of open source software take after data protection issues are reported.

The remainder of this paper is structured as follows: In Section 3, we give a short overview of the related work. Next, we introduce our research questions in Section 4. Then, we describe our research protocol, including materials, variables, execution plan, and analysis plan in Section 5. In Section 6, we describe the limitations and conclude the paper with an overview of the generated data in Section 8.

2 MOTIVATING EXAMPLES

Figures 1 and 2 show examples for two data privacy issues reported on GitHub. The first issue is directly motivated by one of the regulations, i.e., the GDPR, and requests compliance with this legislation. The second issue was not directly motivated by data protection regulations, but still related to an issue in a software. The reporter requests a data anonymization feature to avoid that private data are shared while discussing bugs related to the software product.

While we note that these issues are cherry-picked, they demonstrate the range of activities related to data protection that software developers need to deal with. On one end of the spectrum, we have compliance issues with regulations in software products that requires notices and consent to be correctly implemented. On the other end of the spectrum, we have a general desire to avoid sharing private data when it comes to improving a software product. The spectrum might even be broader, as we have not found related studies analyzing issues on the topic of data protection. The gap that our work plans to address is providing knowledge about how data protection regulations and a general awareness for data protection and privacy impacts software development.

Make the website GDPR-compatible #35

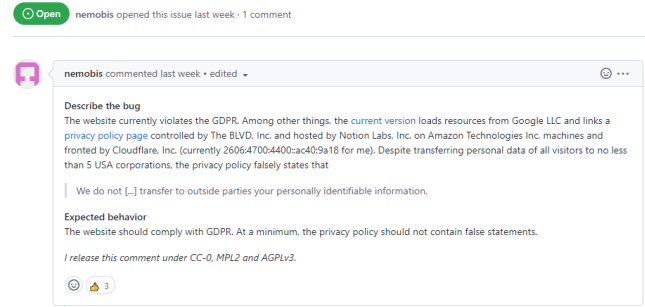


Figure 1: Example for a data protection issue that is reported due to the GDPR (<https://github.com/TheBLVD/mammoth/issues/35>).

Exchange private data inside maps with gibberish data -> for sharing buggy maps without publishing private data #702

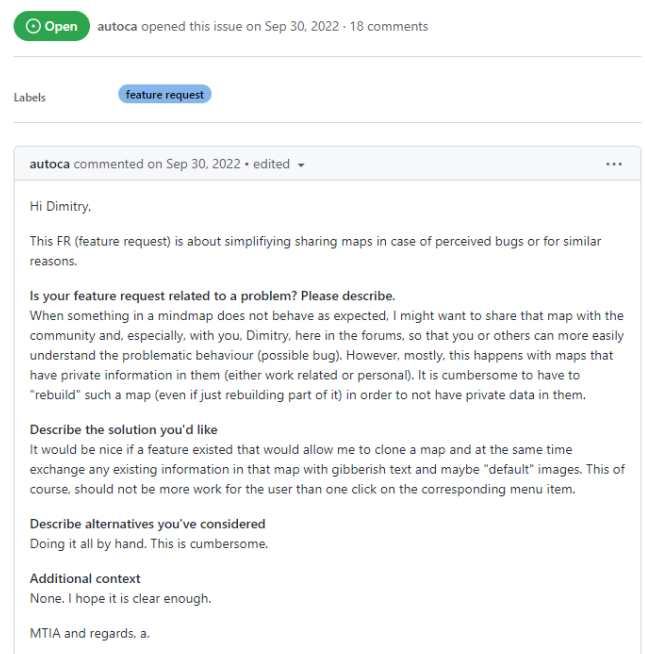


Figure 2: Example for a data protection related issue from the freeplane project that does not specifically mention a legislation (<https://github.com/freeplane/freeplane/issues/702>).

3 RELATED WORK

Due to its huge number of users and projects, GitHub is a common source for historic data about software development, e.g., when studying social aspects of software development (e.g., [23], code reviews (e.g., [43]), or source code history (e.g., [24]). There is also research particular to GitHub issues, e.g., regarding their labels (e.g., [53]) and types (e.g., [29]). Furthermore, there is prior

work that focuses on particular aspects of discussions on GitHub, e.g., security [41], end-user issues [32], or the relation of emotion and issue closing [18]. While not directly related to our work, these papers demonstrate the suitability of GitHub and GitHub issues as study subjects, including the study of how developers deal with specific aspects of software, like security aspects.

Previous research has shown that despite the introduction of stricter data protection laws in recent years, software still fails to comply with the restrictions (e.g., [7, 11, 15, 20, 30, 36, 48]). Therefore, it is an interesting question how data protection topics are discussed among developers. There is, for example, research analyzing discussions on privacy related issues among developers via lab studies [44], in-depth interviews [21, 26, 31, 34, 40] and/or online surveys [8, 33, 45].

While lab studies, interviews and surveys provide important insights into how developers understand privacy and data protection in general, these methods initially stimulated the participants to think about issues related to data protection. Whereas analyses of discussions in various forums are helpful to gain more genuine opinions or understandings without giving a prompt [35].¹ Most notably, three analyses were conducted on understanding personal data and privacy discussions in developer forums [19, 35, 46].

Greene and Shilton [19] analyzed privacy related discussion in an iOS developer forum (iPhoneDevSDK) and an Android forum (XDA) to identify how the term “privacy” is discussed, defined, and framed in contrasting communities. The authors found that privacy is highly influenced by the platform’s philosophy and therefore used fundamentally different in both forums. Rather than talking about “privacy by design”, it should better read “privacy by platform” for the mobile context.

Tahaei et al. [46] used topic modeling techniques to identify privacy-related questions on Stack Overflow and qualitatively analyzed a random sample. The authors identified the following topics which are discussed among developers: privacy policies, privacy concerns, access control, and version changes, with app-related questions being an overarching issue for developers. The authors found that personal concerns as well as client or company requirements were mainly inspiring discussions, whereas laws and regulations, e.g., GDPR related questions, were the least common drivers for discussions. This is also reflected in the work of Bissanyande et al. [10], where “privacy” was not among the top 10 issues reported in a sample of around 20,000 open source projects on GitHub.

Interestingly, Tahaei et al. [46] found that the first questions on “privacy” were created in 2008 (when the StackOverflow was launched), followed by a – more or less – continuous increase in questions over the next 10 years until a sudden decline in 2019. There was no evaluation whether the drivers or the topics have changed over years, so it would be interesting to see, if, e.g., an increase in discussions driven by data protection law correlates with amendments to the law.

Li et al. [35] conducted a qualitative analysis of Reddit posts on issues related to personal data in an Android developer forum. Interestingly, the authors found that most developers rarely discussed privacy concerns during development or implementation of an app,

¹Forums have been used in various contexts to analyze topics that are discussed among developers, e.g. on Stack Overflow [6, 9, 52], or as a comparison between Stack Overflow and GitHub [22, 50].

but rather when the discussion was stimulated by external events, e.g., new privacy laws.

However, all of these analyses are focused mainly on other platforms, like Stack Overflow, and/or the mobile context. To the best of our knowledge, there is no work directly related to ours, i.e., studying issue discussions about personal data and data protection on GitHub for Open Source Software in general. Furthermore, it has not been researched yet, how different roles are interacting in data protection related discussions and what this means for the software development process.

Bissanyande et al. [10] characterized the reporting behavior for open source projects on GitHub in general. The authors found that the majority of projects record a small amount of reported issues. Only about 8% of projects were found to have more than 100 issues reported. Most issues are reported by developers with a large amount of followers and mainly for larger and established projects with popular owners and a large number of watchers and forks. Furthermore, they found that issue reporters, even if they do not belong to the development team, contribute to the code base in most cases. Besides the number of reporters and their contribution to the project, there was no further classification of roles. It will be interesting to see if these findings also hold for issues related to data protection and personal data, or if we find, e.g., more one-time reporters who report the same issue to several software projects.

4 RESEARCH QUESTION

Considering the questions that were left open in the related work, we state the following research questions:

RQ 1: *What kind of issues related to personal data and data protection are discussed on GitHub?*

RQ 2: *How often are those issues related to personal data and data protection reported?*

RQ 3: *Who reports and discusses issues related to personal data and data protection on GitHub?*

RQ 4: *How do developers react to such reported issues?*

Thus, we study the topic from three perspectives. The first perspective is the type of issues that are reported. Based on our current knowledge, the scope of reported issues is unclear. They could, e.g., include data protection related questions regarding project management resources (e.g., mailing lists, or documentation websites), requests to update imported software to a newer version that supports data protection regulations, requests to provide additional features to enable data protection, or requests for the removal of functionality to preserve data protection. The second perspective considers the reporters of the issues. Just like other issues regarding a software, they could come from outside of the development team (e.g., end-users of a software, developers who imported a library) or from inside the development team. The third perspective considers what happens after the reporting, e.g., whether issues are ignored, discussed without resolution, resolved, or rejected.

By considering all three perspectives not only individually, but also combined, we enable a deeper understanding of issues related to personal data and data protection that goes beyond the issues themselves. We can understand if different actors tend to report different issues, if the reactions by developers are different based on

the type of the data protection concern, or if reactions may depend on the role of the reporter.

5 RESEARCH PROTOCOL

We now define the materials, variables, execution plan, and analysis plan of our research protocol. An overview of all steps described in this section can be found in Figure 3.

5.1 Materials

Our study is based on GitHub data from April 2016² until December 2022. Subsequent analysis of the reactions of developers on issues may also involve additional materials with resources from the projects we study, such as mailing lists and externally hosted issue trackers, like Jira or project policies.

5.1.1 Subjects. The subjects of our investigation are GitHub issues that mention or discuss personal data or data protection concerns, and are reported in English. However, we do not consider all of GitHub, but rather limit our investigation to issues from projects that fulfill the following criteria:

- public projects that are not forks and contain a license agreement;
- projects with at least ten contributors after June 2018³;
- projects with active development that have at least 100 commits after June 2018; and
- active usage of GitHub issues with at least 20 issues reported after June 2018.

These criteria guarantee that the projects we consider have at least a small community, active development, and are actively using the reporting mechanism we study, i.e., GitHub issues.

We will apply a keyword search within the issue title and body to determine the data, and retrieve all issues for which the title or text contains one of the terms listed in Table 1. We selected these terms to cover a wide range of terminology associated with personal data and data protection, based on both the goals and aspects of data protection regulations (protection of personal data) as well as their impact (e.g., on cookie usage). We avoided terms like “vulnerability”, “encryption”, or “social networks” because they likely lead to many false positives, even though they are sometimes related to data protection issues. A preliminary search found 21,608 unique issues from 5,892 projects that meet our terms and fulfill the inclusion criteria.⁴

We will then create a random sample of 650 issues for our subsequent analysis. We manually validate for all issues in this sample if they are indeed about personal data and data protection. False positives are removed and we will sample additional issues until we have achieved our desired sample size. The rationale for the sample size of 650 is explained when discussing our methods within the analysis plan (see Section 5.4).

5.2 Variables

We will measure the following variables for each of our subjects.

²The GDPR was adopted April 14th, 2016. Other legislation like the CCPA is younger and, therefore, also covered.

³The GDPR was enforceable starting May 25th, 2018. Other legislation like the CCPA is younger, and there may have been less activity after it’s adoption.

⁴Search conducted between Jan 23rd and Jan 31st, 2023

Terms used to find issues

anonymization, CCPA, consent withdrawal, cookie banner, cookie law, cookie notice, cookie prompt, data breach, data privacy, data protection, data sharing, ePrivacy Directive, fingerprinting, GDPR, personal data, personally identifiable information, PII, privacy act, privacy breach, privacy controls, privacy issue, privacy law, privacy notice, privacy policy, privacy problem, privacy settings, privacy violation, pseudonymization, right to be forgotten, tracking

Table 1: Alphabetical list of the keywords used to identify candidates of issues regarding personal data and data protection.

- *Reporter*: the role of the reporter⁵ of an issue within the project as one of the following: frequent reporter, one-time reporter (only active within a single issue), frequent committer, one-time committer (only active within a single commit or pull request).
- *Discussants*: the role of the discussants⁶ of an issue within the project, analogue to the roles of the reporters.
- *Labels*: the labels assigned to the issue on GitHub, e.g., bug, question, or enhancement. In order to avoid multiple labels with different names, but the same semantics (e.g., bug and defect), we will manually generate a mapping for synonyms similar to the work by Herbold et al. [24], where this was done for Jira issue types.
- *#Comments*: the number of comments in the discussion of the issue.
- *#Discussants*: the number of individuals involved in the discussion.
- *Reporting date*: the date the issue was reported.
- *Last active date*: the date of the last activity related to the issue.
- *Status*: whether the issue is open or closed.
- *Privacy issue*: type of privacy issue.⁷
- *Consent interaction*: interaction to obtain data collection consent
- *Resolution*: actions (if any) taken to address the issue.

The classification into the roles of reporter and discussants is similar to the work by, e.g., Joblin et al. [27], and Honsel et al. [25], who differentiate between core and peripheral developers based on the activity within the project. In our work, we use a simpler approach that uses a very strict definition of peripheral, i.e., only being active once. The intend of this is to identify if somebody was active just because of one data protection issue.

5.3 Execution Plan

Once we have identified the sample of subjects of our study according to the criteria presented in Section 5.1.1, we will automatically collect the data for the variables *Reporter*, *Labels*, *#Comments*, *#Discussants*, *Reporting date*, *Last active date*, and *Status* using the

⁵Person who created an issue

⁶People involved in the discussion of an issue

⁷While we use only the term privacy here, this includes personal data and data protection related issues as well.

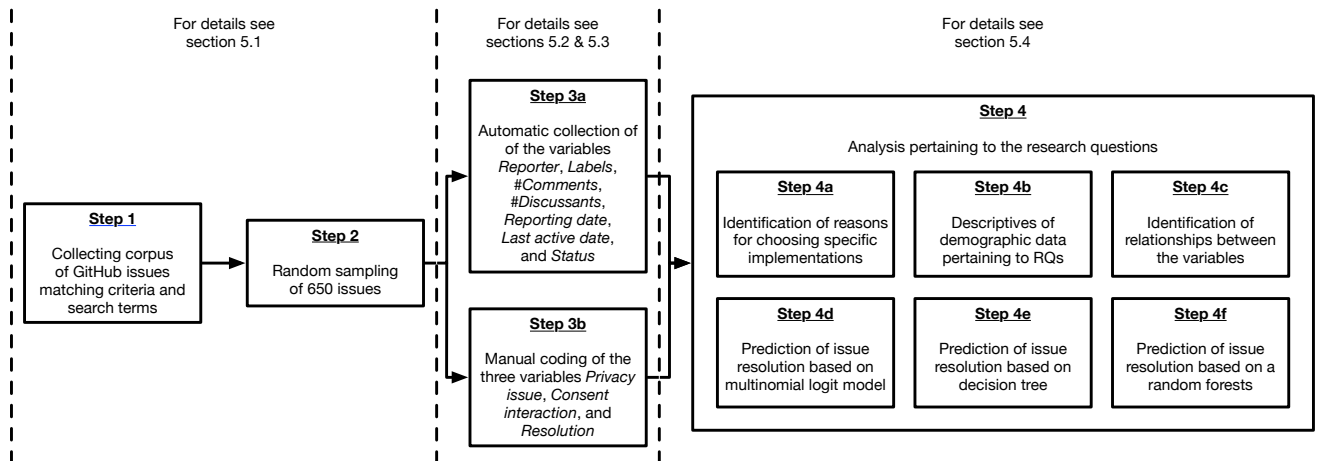


Figure 3: Overview of the steps involved in the methodology of this investigation.

appropriate GitHub APIs. The values for the other three variables *Privacy issue*, *Consent interaction*, and *Resolution* will be obtained through manual coding. As input for the coding, we do not only collect the issues themselves, but also associated pull requests and commits.

The actual manual coding will be conducted using inductive coding [47]. First, two of the authors will independently code the same 20% of the collected data to create the code book. This is common practice in other works in the area of usable security to ensure inter-rater reliability, used e.g. in [37, 39]. It is, furthermore, part of the recommendations by Elder et al. [16]. Cohen’s Kappa will be calculated to measure inter-rater reliability (IRR) of the coding. Not only will IRR be calculated, but the two coders will also meet to discuss their codes and ambiguities in the coding to harmonize the code book they each created up to this point. In the unlikely event that the IRR is below 0.7 after 20% of data have been coded, additional 5% increments of data will be coded to ensure an IRR above 0.7. Once a sufficient IRR above 0.7 has been reached, the remaining data will be coded independently. The coders will, however, continue to discuss ambiguities in the coding, including new codes in the unlikely event that they occur. These methods are deemed sufficient given the exploratory nature of the study.

5.4 Analysis Plan

In the following, we describe our analysis plan. First, we discuss the qualitative analysis we conduct based on what we observe during coding regarding the reasons for resolutions. Then we describe demographics of our population, followed by the quantitative analysis between the variables and the possibility to model predictive models for issue resolutions.

5.4.1 Reasons for Choosing Specific Implementations & Solutions. The qualitative analysis will aim to identify the reasoning for choosing specific implementations for specific issues related to personal data and data protection, and to which resolutions these implementations lead. Additionally, the coders will consider any further aspects of note that come up as part of this exploratory analysis. Count data of all codes will be reported (as opposed to percentages)

to avoid over-generalizing. The individual concepts uncovered in the coding will be illustrated using quotes from the data.

5.4.2 Descriptives of Demographic Data Pertaining to Research Questions. Additionally, we will conduct an empirical analysis based on the data we collect. We will present demographic data about our subjects to determine the prevalence of issues related to personal data and data protection. For this, we will report the number of projects that match our criteria, the number of projects for which we identified at least one issue related to personal data and data protection, the number of overall issues of the projects, and the number of issues related to personal data and data protection identified by the keyword search. Moreover, we will present for each keyword how many of the issues, which are identified using that keyword, had to be discarded during the sampling to measure the false positive rate of the keyword search.

Then, we will consider data for the individual perspectives of our research question. To understand the reporters, we will consider the number of reporters per different type of related issue. Similarly, we consider who joins discussions about issues related to personal data and data protection by reporting the numbers of discussants per different type. Furthermore, we will report these numbers over time, i.e., how many reporters/discussants of each type participated in each issue related to personal data and data protection per quarter within our reporting time frame, as well as the total number of issues related to personal data and data protection within each time frame. To understand what was reported, we will report the results of the manual coding of the issue types, including the description and frequency of issue types related to personal data and data protection. In the same way, we will provide data about the resolutions, i.e., the types of resolutions we observed including a description and the frequencies.

5.4.3 Relationships Between the Variables. In addition to the individual reporting regarding these aspects, we will also evaluate the relationships between our variables. We will evaluate the relationship between the issue types and reporters through the absolute numbers of each reported issue type per reporter type. We will

augment this by an analysis of the cross-tabulation between the nominal reporter type and the numeric issues per type. We will use the χ^2 test with a significance level of $\alpha = 0.05$ to get further information regarding the significance of the relationship. We restrict the statistical test to issue types which we observe at least 20 times, i.e., five times as often as the number of reporter roles.

5.4.4 Prediction of Issue Resolution. Furthermore, we want to understand how the different aspects affect the resolution of issues related to personal data and data protection. For this, we will try to predict the issue resolution as dependent variable based on the other variables as independent variables, i.e., the *Reporter*, *Discussants*, *Labels*, *#Comments*, *#Discussants*, *Status*, *Privacy issue*, and *Consent interaction*. Since a discussion of an issue can have discussants of multiple roles, we encode this variable with four binary variables that mark for each role if there is a discussant with that role in binary form (i.e., 1 if the discussant role is present, 0 if not). Consequently, we have eleven independent variables for these models: two numeric variables (*#Comments* and *#Discussants*) and nine nominal variables (*Reporter*, each of the four *Discussant types*, *Labels*, *Status*, *Privacy issue*, and *Consent interaction*). We take the pattern from Tunkel and Herbold [49] and create multiple models to understand the relationships between our variables: 1) a multinomial logit model to understand the relationship between the independent variables and the odds of the resolution; 2) a decision tree to understand if we can find a description based on Boolean rules for the resolution; and 3) a random forest to understand if a powerful non-linear approach can model the relationship. This multi-perspective approach means that we combine less powerful models that are easy to interpret (linear model for coefficient relationships, rule-based models to understand how concrete values behave) with a more powerful non-linear model to avoid assuming the lack of a relationship as a consequence of underpowered modeling techniques. Based on Bujang et al. [14], we estimate that we require $n = 100 + 50 \cdot \# \text{independent variables} = 100 + 50 \cdot 11 = 650$ issues related to personal data and data protection for the multinomial logit model. In the absence of similar rules for the other models, we use 650 as required sample size for our study.

Prediction Based on Multinomial Logit Model. The multinomial logit model uses a one-hot encoding for the nominal variables *Reporter*, *Status*, *Privacy issue*, and *Consent interaction*. For each variable, we will use Wald's test [51] to determine the significance of each coefficient and interpret the coefficients to understand the impact on the odds of the resolution for each significant variable. Furthermore, we will report McFadden's adjusted R^2 [38] to report the goodness of fit of the model. This will help us to further understand the reliability of the odds, as the coefficients of a model with a poor fit are less reliable.

Prediction Based on Decision Tree. The decision tree can directly work with the nominal data and does not require one-hot encoding. We will use a CART decision tree [13] with Gini impurity as splitting criterion. The choice of splitting criterion has been shown to not have a large impact on the resulting trees (see, e.g., [42]). We will not restrict the tree depth and conduct a manual analysis of the resulting decision tree. Thus, instead of using the overall accuracy to determine the quality of the model which may have problems with

overfitting, we will rather consider the individual data partitions at the nodes of the decision trees, as this allows us better and more fine-grained insights. We will consider which decision were made, how the decisions help to decide for specific resolutions, as well as the general support of the decisions, i.e., the amount of data used for the decision and within the resulting subsets.

Prediction Based on Random forests. Random forests [12] are consistently among the best performing machine learning models for smaller tabular data sets [17]. A random forest determines a non-linear relationship between the dependent variable and the independent variable through an ensemble of decision trees, where each decision tree is trained on a subset of the data and variables. In contrast to the decision tree, we cannot feasibly manually analyze a decision tree to understand the relationship, as we would have to consider hundreds of trees. Instead, we will use this analysis to augment our insights from the less powerful but interpretable decision trees. Concretely, we will calculate the feature importance, which measures how much each feature contributed to the reduction of the Gini impurity that is observed at the leaf nodes of the trees, averaged over all trees.

We will augment the above analyses for the relation of the resolution with an analysis of the confusion matrices, to understand if the models are better at modeling some resolutions than others. Furthermore, we compute the correlations with Spearman's ρ between all variables, as this allows us to understand interactions between variables within the models.

6 LIMITATIONS

There are several limitations of our work. While we try to determine a large and unbiased samples of issues related to personal data and data protection from GitHub, our data collection approach may still introduce some biases. Our criteria for projects exclude small projects with few contributors or general development activity since June 2018. However, we note that activity and contributor based filtering was identified as a suitable strategy to avoid problems when analyzing data from GitHub [28]. Moreover, our search-based approach based on a list of terms could possibly miss issues related to personal data and data protection, in case none of the identified terms is mentioned as is, e.g., because we missed terms or due to typos. Due to this, we cannot rule out that we will miss types of issues related to personal data and data protection that are not captured by our search.

Moreover, since our study is restricted to GitHub as a data source, we cannot generalize our conclusion to reporting of issues related to personal data and data protection in general. For example, users not familiar with software development may not be aware of GitHub and would contact developers in a different way, e.g., social networks or mailing list. Our study is not suitable to capture such issues, unless the developers would then create an issue on GitHub related to this. In extension to this, we also only capture issues created (and discussed) in English. While we believe that the majority of open source projects that are relevant for our research falls into this category, projects where discussions happen in different languages might offer additional insights. Including these, however, is beyond the scope of this paper.

We only consider GDPR (EU), DPA (UK), PA (CAN), and CCPA (CA, USA) as regulations in our keywords. Including regulations from additional jurisdictions might have broadened our search, but was also beyond the scope of this paper.

7 PUBLICATION OF GENERATED DATA

All code and data will be made publicly available following the FAIR principles. The code for data collection and analysis will be made available on GitHub and, additionally, stored in a DOI citable long-term archive like Zenodo. Depending on the size of the data, the data will either be shared together with the code or within a separate DOI citable long-term archive. The shared data will at least include the following.

- Copies of the studied issues, including their discussion. We will sanitize these copies to remove personal data, such as email addresses or usernames, using regular expressions.
- Measurements of all variables for the studied issues.
- Data about the agreement for the manual coding.

8 CONCLUSION

Within this study, we want to understand the reporting and resolution of issues related to personal data and data protection on GitHub. Previous studies have investigated different topics or discussions about personal data and data protection in different channels. To the best of our knowledge, no study has yet researched discussions about personal data and data protection on GitHub. Furthermore, related work either focused on discussions that were started by developers themselves, or discussions that were more general and not tied to a certain project. Thus, within this exploratory study, we want to shed light on the impact of data protection regulations throughout the whole process of the software development on GitHub. We study this with a combination of qualitative and quantitative analysis to understand what is reported, who reports and resolves issues, and how the different aspects are correlated to each other. It will be interesting to see, whether we can confirm prior work, e.g. with respect to the topics discussed, or the drivers that spark discussions. We will face a number of limitations, but as this study will be the first of its kind, we hope that future studies will continue our work.

REFERENCES

- [1] [n. d.]. California Consumer Privacy Act (CCPA). <https://www.oag.ca.gov/privacy/ccpa>
- [2] [n. d.]. DIRECTIVE 2002/58/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32002L0058&qid=1674304867887&from=EN>
- [3] [n. d.]. REGULATION (EU) 2016/ 679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL - of 27 April 2016 - on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/ 46/ EC (General Data Protection Regulation). <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>
- [4] [n. d.]. Unwrapping the consent box. The CJEU Judgment in the Orange Romania Case – European Law Blog. <https://europeanlawblog.eu/2020/12/10/unwrapping-the-consent-box-the-cjeu-judgment-in-the-orange-romania-case/>
- [5] [n. d.]. The “Planet49” decision of the German Federal Court of Justice – ePrivacy. <https://www.eprivacy.eu/en/news/news-detail/news/die-planet49-entscheidung-des-bgh/>
- [6] Rabe Abdalkareem, Emad Shihab, and Juergen Rilling. 2017. What Do Developers Use the Crowd For? A Study Using Stack Overflow. *IEEE Softw.* 34, 2 (mar 2017), 53–60. <https://doi.org/10.1109/MS.2017.31>
- [7] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions Speak Louder than Words: Entity-Sensitive Privacy Policy and Data Flow Analysis with POLICHECK. In *Proceedings of the 29th USENIX Conference on Security Symposium (SEC'20)*. USENIX Association, USA, Article 56, 18 pages.
- [8] Rebecca Balebako, Abigail Marsh, Jialiu Lin, Jason I. Hong, and Lorrie Cranor. 2014. The Privacy and Security Behaviors of Smartphone App Developers. (2 2014). <https://doi.org/10.1184/R1/6470528.v1>
- [9] Anton Barua, Stephen W. Thomas, and Ahmed E. Hassan. 2014. What are developers talking about? An analysis of topics and trends in Stack Overflow. *Empirical Software Engineering* 19, 3 (2014), 619–654. <https://doi.org/10.1007/s10664-012-9231-y>
- [10] Tegawendé F. Bissyandé, David Lo, Lingxiao Jiang, Laurent Réveillère, Jacques Klein, and Yves Le Traon. 2013. Got Issues? Who Cares About It? A Large Scale Investigation of Issue Trackers from GitHub. *2013 IEEE 24th International Symposium on Software Reliability Engineering (ISSRE)* (2013), 188–197. <https://doi.org/10.1109/issre.2013.6698918>
- [11] Dino Bollinger, Karel Kubicek, Carlos Cotrini, and David Basin. 2022. Automating Cookie Consent and GDPR Violation Detection. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 2893–2910. <https://www.usenix.org/conference/usenixsecurity22/presentation/bollinger>
- [12] Leo Breiman. 2001. Random Forests. *Mach. Learn.* 45, 1 (Oct. 2001), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [13] L. Breiman, J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.
- [14] Mohamad Adam Bujang, , Nadiah Sa'at, Tg Mohd Ikhwan Tg Abu Bakar Sidik, Lim Chien Joo, , and and. 2018. Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. *Malaysian Journal of Medical Sciences* 25, 4 (2018), 122–130. <https://doi.org/10.21315/mjms2018.25.4.12>
- [15] Saksham Chitkara, Nishad Gothoskar, Suhars Harish, Jason I. Hong, and Yuvraj Agarwal. 2017. Does This App Really Need My Location? Context-Aware Privacy Management for Smartphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 42 (sep 2017), 22 pages. <https://doi.org/10.1145/3132029>
- [16] Glen H Elder, Eliza K Pavalko, and Elizabeth Colerick Clipp. 1993. *Working with archival data: Studying lives*. Vol. 88. Sage.
- [17] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinami Amorim. 2014. Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *J. Mach. Learn. Res.* 15, 1 (Jan. 2014), 3133–3181.
- [18] I. Ferreira, B. Adams, and J. Cheng. 2022. How heated is it? Understanding GitHub locked issues. In *2022 IEEE/ACM 19th International Conference on Mining Software Repositories (MSR)*. IEEE Computer Society, Los Alamitos, CA, USA, 309–320. <https://doi.org/10.1145/3524842.3527957>
- [19] Daniel Greene and Katie Shilton. 2018. Platform privacies: Governance, collaboration, and the different meanings of “privacy” in iOS and Android development. *New Media & Society* 20, 4 (2018), 1640–1657. <https://doi.org/10.1177/1461444817702397>
- [20] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. 2020. “It’s a Scavenger Hunt”: Usability of Websites’ Opt-Out and Data Deletion Choices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376511>
- [21] Irit Hadar, Tomer Hasson, Oshrat Ayalon, Eran Toch, Michael Birnhack, Sofia Sherman, and Arod Balissa. 2018. Privacy by designers: software developers’ privacy mindset. *Empirical Software Engineering* 23, 1 (2018), 259–289. <https://doi.org/10.1007/s10664-017-9517-1>
- [22] Junxiao Han, Emad Shihab, Zhiyuan Wan, Shuiguang Deng, and Xin Xia. 2020. What do Programmers Discuss about Deep Learning Frameworks. *Empirical Software Engineering* 25, 4 (2020), 2694–2747. <https://doi.org/10.1007/s10664-020-09819-6>
- [23] Steffen Herbold, Aynur Amirfalah, Fabian Trautsch, and Jens Grabowski. 2021. A systematic mapping study of developer social network research. *Journal of Systems and Software* 171 (2021), 110802. <https://doi.org/10.1016/j.jss.2020.110802>
- [24] Steffen Herbold, Alexander Trautsch, Fabian Trautsch, and Benjamin Ledel. 2022. Problems with SZZ and features: An empirical study of the state of practice of defect prediction data collection. *Empirical Software Engineering* 27, 2 (Jan. 2022). <https://doi.org/10.1007/s10664-021-10092-4>
- [25] Verena Honsel, Steffen Herbold, and Jens Grabowski. 2016. Hidden Markov Models for the Prediction of Developer Involvement Dynamics and Workload. In *Proceedings of the The 12th International Conference on Predictive Models and Data Analytics in Software Engineering* (Ciudad Real, Spain) (PROMISE 2016). Association for Computing Machinery, New York, NY, USA, Article 8, 10 pages. <https://doi.org/10.1145/2972958.2972960>
- [26] Leonardo Horn Iwaya, Muhammad Ali Babar, and Awais Rashid. 2022. Privacy Engineering in the Wild: Understanding the Practitioners’ Mindset, Organisational Culture, and Current Practices. <https://doi.org/10.48550/ARXIV.2211.08916>
- [27] Mitchell Joblin, Sven Apel, Claus Hunsen, and Wolfgang Mauerer. 2017. Classifying developers into core and peripheral: An empirical study on count and

- network metrics. In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*. IEEE, 164–174.
- [28] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M. German, and Daniela Damian. 2015. An in-depth study of the promises and perils of mining GitHub. *Empirical Software Engineering* 21, 5 (Sept. 2015), 2035–2071. <https://doi.org/10.1007/s10664-015-9393-5>
- [29] Rafael Kallis, Oscar Chaparro, Andrea Di Sorbo, and Sebastiano Panichella. 2022. NLBSE'22 Tool Competition. In *2022 IEEE/ACM 1st International Workshop on Natural Language-Based Software Engineering (NLBSE)*. 25–28. <https://doi.org/10.1145/3528588.3528664>
- [30] Georgios Kampanos and Siamak F. Shahandashti. 2021. Accept All: The Landscape of Cookie Banners in Greece and the UK. In *ICT Systems Security and Privacy Protection*, Audun Jøsang, Lynn Fletcher, and Janne Hagen (Eds.). Springer International Publishing, Cham, 213–227.
- [31] D. Kekulluoglu and Y. Acar. 2023. "We are a startup to the core": A qualitative interview study on the security and privacy development practices in Turkish software startups. In *2023 2023 IEEE Symposium on Security and Privacy (SP) (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 1331–1347. <https://doi.org/10.1109/SP46215.2023.00076>
- [32] Hourieh Khalajzadeh, Mojtaba Shahin, Humphrey O. Obie, and John Grundy. 2022. How are Diverse End-user Human-centric Issues Discussed on GitHub?. In *2022 IEEE/ACM 44th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. 79–89. <https://doi.org/10.1145/3510458.3513014>
- [33] Patrick Kührtreiber, Viktoriya Pak, and Delphine Reinhardt. 2022. A survey on solutions to support developers in privacy-preserving IoT development. *Pervasive and Mobile Computing* 85 (2022), 101656. <https://doi.org/10.1016/j.pmcj.2022.101656>
- [34] Tianshi Li, Yuvraj Agarwal, and Jason I. Hong. 2018. Coconut: An IDE Plugin for Developing Privacy-Friendly Apps. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 178 (dec 2018), 35 pages. <https://doi.org/10.1145/3287056>
- [35] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I. Hong. 2021. How Developers Talk About Personal Data and What It Means for User Privacy. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28. <https://doi.org/10.1145/3432919>
- [36] Célestin Matte, Nataliia Bielova, and Cristiana Santos. 2020. Do Cookie Banners Respect my Choice?: Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. In *2020 IEEE Symposium on Security and Privacy (SP)*. 791–809. <https://doi.org/10.1109/SP40000.2020.00076>
- [37] Peter Mayer, Yixin Zou, Florian Schaub, and Adam J. Aviv. 2021. "Now I'm a bit angry." Individuals' Awareness, Perception, and Responses to Data Breaches that Affected Them. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security Symposium)*. USENIX Association, 393–410. <https://www.usenix.org/conference/usenixsecurity21/presentation/mayer>
- [38] D. McFadden. 1974. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics* (1974), 105–142.
- [39] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2019. Why people (don't) use password managers effectively. In *Symposium on Usable Privacy and Security (Symposium on Usable Privacy and Security)*. 319–338. <https://www.usenix.org/conference/soups2019/presentation/pearman>
- [40] Mariana Peixoto, Dayse Ferreira, Mateus Cavalcanti, Carla Silva, Jéssyka Vilela, João Araújo, and Tony Gorschek. 2023. The perspective of Brazilian software developers on data privacy. *Journal of Systems and Software* 195 (2023), 111523. <https://doi.org/10.1016/j.jss.2022.111523>
- [41] Daniel Pletea, Bogdan Vasilescu, and Alexander Serebrenik. 2014. Security and Emotion: Sentiment Analysis of Security Discussions on GitHub. In *Proceedings of the 11th Working Conference on Mining Software Repositories (Hyderabad, India) (MSR 2014)*. Association for Computing Machinery, New York, NY, USA, 348–351. <https://doi.org/10.1145/2597073.2597117>
- [42] Laura Elena Raileanu and Kilian Stoffel. 2004. Theoretical Comparison between the Gini Index and Information Gain Criteria. *Annals of Mathematics and Artificial Intelligence* 41, 1 (May 2004), 77–93. <https://doi.org/10.1023/b:amai.0000018580.96245.c6>
- [43] N. Rao, J. Tsay, M. Hirzel, and V. J. Hellendoorn. 2022. Comments on Comments: Where Code Review and Documentation Meet. In *2022 IEEE/ACM 19th International Conference on Mining Software Repositories (MSR)*. IEEE Computer Society, Los Alamitos, CA, USA, 18–22. <https://doi.org/10.1145/3524842.3528475>
- [44] Awanthika Senarath and Nalin A. G. Arachchilage. 2018. Why Developers Cannot Embed Privacy into Software Systems? An Empirical Investigation. In *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018 (Christchurch, New Zealand) (EASE'18)*. Association for Computing Machinery, New York, NY, USA, 211–216. <https://doi.org/10.1145/3210459.3210484>
- [45] Swapneel Sheth, Gail Kaiser, and Walid Maalej. 2014. Us and Them: A Study of Privacy Requirements across North America, Asia, and Europe. In *Proceedings of the 36th International Conference on Software Engineering (Hyderabad, India) (ICSE 2014)*. Association for Computing Machinery, New York, NY, USA, 859–870. <https://doi.org/10.1145/2568225.2568244>
- [46] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. 2020. Understanding Privacy-Related Questions on Stack Overflow. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3313831.3376768>
- [47] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (2006), 237–246. <https://doi.org/10.1177/1098214005283748>
- [48] Martino Trevisan, Stefano Traverso, Eleonora Bassi, and Marco Mellia. 2019. 4 Years of EU Cookie Law: Results and Lessons Learned. *Proceedings on Privacy Enhancing Technologies* 2019, 2 (2019), 126–145. <https://doi.org/10.2478/popets-2019-0023>
- [49] Steffen Tunkel and Steffen Herbold. 2022. Exploring the relationship between performance metrics and cost saving potential of defect prediction models. *Empirical Software Engineering* 27, 7 (Sept. 2022). <https://doi.org/10.1007/s10664-022-10224-4>
- [50] Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. 2013. StackOverflow and GitHub: Associations Between Software Development and Crowdsourced Knowledge. *2013 International Conference on Social Computing* (2013), 188–195. <https://doi.org/10.1109/socialcom.2013.35>
- [51] Abraham Wald. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* 54, 3 (1943), 426–482. <https://doi.org/10.1090/s0002-9947-1943-0012401-3>
- [52] Yuhao Wu, Shaowei Wang, Cor-Paul Bezemer, and Katsuro Inoue. 2019. How do developers utilize source code from stack overflow? *Empirical Software Engineering* 24, 2 (2019), 637–673. <https://doi.org/10.1007/s10664-018-9634-5>
- [53] Xiaoyuan Xie, Yuhui Su, Songqiang Chen, Lin Chen, Jifeng Xuan, and Baowen Xu. 2021. MULA: A just-in-time multi-labeling system for issue reports. *IEEE Transactions on Reliability* 71, 1 (2021), 250–263.