Questions of Privacy: An examination of developer inquiries towards the matters of privacy in the last five years

| Anthony Jackson  *Computer Science Department, University of Maine Orono, ME, 04469, USA*  anthony.a.jackson@maine.edu | Christan Doiron  *Computer Science Department, University of Maine Orono, ME, 04469, US*  christian.doiron@maine.edu | Elijah Daran  *Computer Science Department, University of Maine Orono, ME, 04469, US*  elijah.daran@maine.edu |
| --- | --- | --- |

***Abstract*** - **Stack Overflow is a questions and answer website created in early 2008 to act as a place for amateur and professional programmers, designers and analysts to seek help with concepts and practices related to computer science and programming. In recent years this database has often been used for questions related to privacy and security from both a technical and legal perspective. In this paper, we develop an approach for extracting privacy and security questions from Stack Overflow for the purposes of analysis in an effort to understand the common question trends of developers, and to evaluate the effectiveness of Stack Overflow in answering these questions. We will “scrub” the database using an SQL data explorer to catalog one hundred questions within the last five years to uncover common trends and popular advice related to privacy and security policies and practices. Our goal is to understand the nature of the developers and see how their questions and answers might have changed pre, post, and during the enactment of CCPA and GDPR. Our results found that developers often face confusion with regard to how particular libraries or services interact with laws, how privacy policies are created for applications and how to create code which defines safety and security for users. Thus, we postulate that there should be an in website based catalog of application guidelines which detail how software and libraries interact with the law and how answers should be formatted for users to easily comprehend and apply.**

***Keywords -* Stack Overflow, Privacy, Security, Development, Database, Privacy Regulations**

I. Introduction

The concepts of privacy and security have always been a point of contention for developers across the years. This contention is sparked by regulations established by law and often takes the form of arguments about how to best follow its guidelines. These strict regulations often differ from country to country, leaving developers confused on how to apply their specific methodology to the general concept of security and privacy presented within this legal framework. Thus, many of these developers turn to scouring online sources and asking questions on public forums hoping to receive guidance. One of the most popular such forums is known as Stack Overflow, a market of developer-focused information that has “19 million users who’ve asked over 23 million questions” [1].

Within these 23 million questions are numerous focused on privacy and security, with many questions being near-duplicates of each other. Given the amount of overlap between these questions, is there a way to better prepare these developers with materials that can answer the questions without needing to approach the public forum space? In addition, have these questions changed or stayed the same after the creation and implementation of new privacy regulations, such as the General Data Protection Regulation (GDPR) [17] and the California Consumer Privacy Act (CCPA) [18]?

The space that Stack Overflow occupies in the modern software development community has caused it to be the focus of several academic studies. Some have even examined privacy-related questions on Stack Overflow in a very similar manner in order to analyze the modern developer’s mindset when it comes to privacy and security [1]. The relevant papers and articles range from analyses of Stack Overflow questions [5] [13] to studies on modern opinions and sentiments on privacy [4] [8] . Additionally, there are accounts on Stack Overflow’s status as a forum and collection of data itself. All of these papers have their own uses as examples for our work, providing meaningfully context and examples of methods which could either be detrimental or influential in our research and understanding.

In this paper, we describe a pre-emptive solution which can appeal to developers with questions and concerns about privacy features and laws. The ideal solution should lead to more educated, informed, and privacy-minded developers. It should also be relevant to the average person who is likely to consult Stack Overflow when confronted with a potential issue. To ensure this, we sorted through thousands of questions from the website over the past five years and analyzed the common factors in questions relating to privacy. This approach was facilitated by Stack Overflow’s built-in tool to search and sort archived questions and answers.

Incorporating our heavily privacy-focused study with the findings from previous works allowed us to identify key issues that new developers struggled with and propose a solution database hosted by Stack Overflow for the purpose of answering common or core questions often related to the fundamentals of privacy.

The rest of this paper is structured as follows: In Section 2 we describe other paperwork and studies related to the topics of privacy. In Section 3 are some of the solutions, contributions and research methods which we bring to this discussion and research. Section 4 contains the final results of our research, analysis, and conclusions.

II. Relevant Papers.

In Tahaei et al. [2], the researchers aimed to examine the modern state of privacy engineering. They first interviewed 12 advocates for privacy engineering and 20 computer science students and then conducted a survey of 400 mobile developers regarding privacy practices. Finally, they analyzed thousands of questions from Stack Overflow to determine some of the most common ways to solve privacy engineering problems. The result shows that modern privacy problems could be partially fixed by creating standardized privacy regulations that are simple to understand and put into practice. The benefits of this study is the range that the researchers were able to glean information from- not only because of the large number of participants but also because the interviewed developers specialized in several different types of development. One potential shortcoming is that interviewed developers may have lacked experience with serious privacy issues, such as data breaches or leaks. This paper has taught us to view a wide range of issues related to privacy and security on Stack Overflow, approaching our search with a more open understanding that the differences between something like language privacy and application privacy are both equally important.

Acar et al. [3] conducted a survey with 295 developers who had published applications to the Google Play Store as well as a further study involving 54 Android developers, who were given access to several different types of resources and asked to complete programming tasks. The result shows that the group which only had access to Stack Overflow provided code supplements that were far less secure that the other groups code, and that an over-reliance on Stack Overflow leads to less secure code. Our study provides a result which helps bring focus to this statement through our discovery of how developers on Stack Overflow manage their code and how faulty or functional code is spread by Stack Overflow’s user base.

Tahaei et al. [4] conducted a study on all. Stack Exchange posts between July 2016 and July 2022. The researchers looked for questions that had combinations of certain tags related to mobile health apps and analyzed the trends over time. The results of this study indicated that new developers had trouble with specific privacy requirements for their respected platform, not being able to identify what or what was not needed for these applications. This method undertaken in this paper is similar to our own, although we have expanded on this method with new privacy-related keywords and tags to find a more generalized viewpoint for information. Overall, this study is incredibly useful and was taken under consideration as we conducted our own research and challenged the problem.

In Allamanis et al. [5], the researchers do not spend much time discussing the results of its Stack Overflow analysis that was conducted. Instead, the benefit of this paper is the techniques behind the analysis itself, which utilized the Stack Overflow API in a creative way that allowed the researcher to categorize each question in a specific type for organizational purposes. While this approach was useful for categorization, our methodology expanded on the practices applied in this study to find specific privacy and security related topics that allowed us to draw accurate conclusions.

Ferreyra et al. [6] is a unique paper in that, instead of just analyzing Stack Overflow’s questions and answers, the researchers analyzed the profiles of users behind the questions. The purpose of this study was to determine the level of “self-disclosure” of users who tended to answer more privacy and security related questions. This study found that users who are more active in this privacy and security related topics tended to disclose less information about themselves through their user profiles, suggesting they have a higher standard for their own anonymity. These insights were useful in assessing user accounts and understanding if these answer providers are active members of their communities with personal motivations, or are just users who faced similar concerns and decided to provide answers to others once they understood.

Scoccia et al. [7] is another paper with a very similar methodology to our own. The most obvious difference is the amount of Stack Overflow questions analyzed. The researchers analyzed 10,822 questions and expanded their study to include Github issue reports. The most interesting aspect of this study is the way researchers considered the different skills that were necessary to develop desktop applications, something we could apply to the analysis of our very own data set. The way that these question trends were documented over year-long periods is a method we also utilized, though we limited ourselves to only five years to keep the data relevant and simple.

Rather than analyzing privacy questions that had already been posted and vetted from online forums, Tahaei et al. [8] conducted an online survey which had 400 individual answers, each of them familiar with app development. The study found that ad networks that present choices to developers in a way that highlights the privacy consequences of ad personalization are significantly more likely to choose non-personalized ads for their users. However, this survey could be skewed, as these developers were often individuals who already had some experience with ad networks and weren’t as eager to just gather a quick paycheck like some new app developers might. This study doesn’t have any obvious usage in our discussion, but acts as a good background for understanding the mobile development side of the question and answer spectrum.

Li et al. [9] aimed to examine how Android developers discuss personal data user and privacy on the /r/androiddev “sub-reddit” style forum on Reddit. The study used bottom-up open coding on a sample of 207 threads with a total of 4772 individual posts to develop a typological discussion of personal data use. The results showed that Android developers rarely mentioned privacy concerns when discussing specific application design or implementation problems. However, they often had active conversations around privacy when stimulated by certain external events which represented new privacy-enchanting restrictions from the Android operating system, application store policies, privacy and/or security laws. Developers often felt that these restrictions could have a considerable cost, yet also failed to generate any compelling benefit for themselves. Given these results, the study provides a set of suggestions for Android OS and the Google Play Store to implement privacy-focused features and for developer forums to encourage more in-depth privacy discussions. While this study bears some differences to ours in both forum medium and amount of data processed, it provides some expectation as to how developers in Stack Overflow discussions might react to these GDPR and CCPA regulations. Our methodology, however, differs greatly- we look towards how these attitudes have changed over time rather than the generalized responses.

Kubicek et al. [10] is a masters thesis which analyzes Github in order to gauge how Github projects have changed in relation to new data laws such as the GDPR, CCPA and CPRA. The approach for this thesis was to extract numerous amounts of metadata, then graph the interactions which mention the key terms to see how it has changed over a period of time. This approach was incredibly sophisticated, although it would obviously be skewed as search vectors would be mentioned up to an extent then normalized until no longer bearing direct mention. In addition, this approach might not find changes which are made without direct reference to the data laws through description or title. While our topic is different in that it revolves around Stack Overflow questions, this study serves as a wonderful example for drawing a direct comparison between results and how ours have come up short or are skewed in unexpected directions.

Mondal et al. [11] sought to determine if the reproducibility of coding issues included in Stack Overflow questions significantly impacted the quality and speed of the responses to said questions received. The study also explored how to measure “reproducibility” and what questions contained non-reproducible code. It then investigated what other factors could influence the speed and quality of answers besides reproducibility of programming-related issues. The study found that questions which had reproducible answers were twice as likely to be accepted, and that the median time for these questions was half as long. This study has several interesting implications, primarily because of their extremely laborious and long form data collection and discussion period. While this example does hold some value in how to classify questions and understand how provided materials from a question asker could be dissected, their entire methodology is far too extensive to be mimicked in the brief window we worked underneath.

Tahaei et al. [12] is important to demonstrate the impact that Stack Overflow has on modern development. The researchers conducted a small-scale survey of 289 projects (30.5% of which contained code from Stack Overflow) as well as a survey of over one million applications on the Google Play store (15.4% of which contained Stack Overflow code). It was also found that an alarming number of these projects contained insecure code. They then analyzed advice given to Stack Overflow questions about privacy- it was found that much of this advice took a quasi-documentational approach, with answers reaching beyond the scope of official guides and documentation. The paper also provided more insight into the motivation of the people who answer the questions that we analyzed in our study, allowing us to have a better understanding of the reason behind their input.

In Tahaei et al.[13], researchers collected 1,733 privacy-related questions and automatically categorized them by topic. A random sample of 315 questions were then qualitatively analyzed, and it was found that a large amount of answers did not reference external resources. The main motivation behind the paper was to determine how developers respond to strict privacy policies enacted by major tech companies. This paper is an extremely valuable resource as it follows a very similar methodology to our plans, although at a more grand scale with a longer collection period. The methodology in term related searching is the same fundamental framework as our own, although theirs was more particular on sub categorical analysis which was not relevant to our current research.

III. Solutions and Contributions.

A. *Main Proposition*.

The main problem that was evident is a fundamental misunderstanding of the requirements of the most common and necessary privacy regulations. Some of the users asking questions seemed to approach their problems from a perspective where they assumed that compliance with these regulations were automatic or would be fixed by a third party. While another common theme is related to questions asking if the usage of certain software would ensure compliance with certain laws- unaware that the responsibility falls on them to utilize said software in a secure manner. Conversely, some developers were excessively concerned about complying with privacy laws that did not apply to them - for example, GDPR requirements for an entirely US-based company and product. Most of these questions were answered with reasonable, clear explanations- from developers who were experienced with a particular piece of software and/or certain privacy regulations. However, it is worth considering the large number of developers at the same level of privacy understanding who did not utilize Stack Overflow or a similar resource. For those developers, and developers who are confused about privacy practices in general, we believe a consolidated, easily digestible database of various privacy software, regulations, and solutions should be created.

This hypothetical database would incorporate several common ideas and sentiments found in the papers we analyzed. Researchers generally agreed that a low level of privacy-related knowledge among developers leads to software that is less secure. Additionally, it was found that it is difficult for developers to adapt to new platforms and the different privacy conventions that came with them. An accessible but detailed catalog- with guides for different platforms and their associated laws- could fulfill the same purpose as Stack Overflow for many new developers without any inconsistency caused by faulty answers or improperly phrased questions. The database would directly benefit from the question data we gathered, as we have identified major themes between the most common kinds of questions asked. Additionally, a brief timeline of relevant privacy laws for major regions could be provided. This would be prudent considering that we only analyzed questions asked after both GDPR and CCPA were implemented, but found many questions about both as if they were unfamiliar.

B. *Further Requirements.*

Utilizing a hypothetical “privacy database” would undoubtedly have a positive impact on the software development community, especially for newer developers. However, it is unclear how effectively or quickly an entirely new website would be shared, and a separate application could not be considered a “solution” to Stack Overflow’s core issues. Therefore, we propose that the catalog be implemented directly into Stack Overflow, providing a simple and effective resource for other users to point curious developers toward. The catalog could also be automatically recommended to users asking questions with privacy-related keywords. However, the catalog would not be a replacement for experienced developers answering more complicated questions- rather, it would be a way for Stack Overflow to pre-emptively help hundreds of developers with privacy-focused questions each year. This would not be a simple undertaking for Stack Overflow- but it could bolster their reputation and guide new developers into a more privacy-focused mindset on and after their first few projects.

IV. Methodology and Implementation

Our methodology began with taking the questions sorted within our Excel sheet and writing a one to two sentence summarization for each of them. This allowed us to succinctly identify each answer and question so that we could properly categorize the problems presented. We cataloged these using simple one to two word identifiers which indicated that they were either related to programming, specific to a software or library, about legality for laws and regulations or only related to privacy policies conceptually. These specifications were chosen after reading through the questions and finding patterns to particular subject matters and how they were answered. Answers which involved large sections of code would be identified as “programming”, unless they are entirely dependent on a single software or library- which was an independent category. The “privacy policy” identifier refers to questions related solely on how to write, present or validate policies. Finally, questions which related to law or bills would be summarized under the category of “legality”. This organized and categorized data could then be used to find common trends to develop our understanding of the problem and define solutions as stated in Part III, Section A.

A. *Toolset.*

Our approach to this study revolved around collecting 100 questions from Stack Overflow-with the only criteria being that they needed to be relevant to privacy or security and be asked or updated between January 1st 2022 and January 1st 2017. While questions after January 1st 2021 were gathered manually, those asked prior were scrubbed using the StackExchange Data Explorer. This SQL database allows users to search popular websites such as Stack Overflow for information related to questions and answers- such as a post identification number, keywords, users and even dates. In our case we sorted this database for posts that were after 2016, contained a keyword phrase, met a threshold for engagement, and a minimum of 500 views. Utilizing the SQL explorer was relatively simple; we used basic database manipulation to force all titles into lowercase as to avoid any capitalization issues, removing dates that were not in the range and ordering them by either user score or views. The posts which were collected from 2022 were done manually with a minimum of 50 views, as the database didn’t correctly represent some 2022 entries and many relevant 2022 posts did not have as many views or user interaction. This is due to the fact that Stack Overflow works off of an engagement system that shows posts with higher interaction first, sometimes causing questions to be filtered into obscurity. All retrieved entries were then placed into an Excel spreadsheet, which automatically sorted the data and provided us an area to analyze the questions and summarize each one into an overlying category depending on what was asked and answered.

B. *Approach Validation.*

Our approach differs from most other sources of inquiry towards Stack Overflow, such as Tahaei, et al. [7] and [12] in that both of these sources utilized a thousand or more questions as a dataset while ours only uses one hundred. While this is a smaller dataset, it is much more manageable, as it allows our three person team to properly analyze each question and assess a small snapshot into the developer world. This means our approach is a process design that could be easily upscaled with more researchers to validate even more content. The approach also falls within guidelines for concurrent validation as researchers verify each other's work to check for inconsistencies during the analysis process as questions are reread and reevaluated, keeping data clean and accurate. Our choice in time period for these questions is also beneficial as it is just within the time frame of both the GDPR and CCPA, two of the most important modern policies related to privacy-serving as the backbone for many regulations in Europe and America. The major function of our approach is the “terminology search method” which is highly adaptable; being easily adjustable for any new bills, laws or additional nomenclature that need to be added to the research pool. This also allows us to use a blacklist to avoid certain terminology which might not be relevant to security. For example, developers seeking assistance with specific services like Firebase or languages like C++ may have code using the word “private” or “privacy” in reference to function declaration rather than actual user or content privacy. This stops data from being artificially skewed by information which isn’t relevant to the research topic and could be used for more specific queries such as research into a specific law, field of study or topic.

V. Results.

Through the process of compiling our database, we discovered some interesting statistics pertaining to the nature of the 100 questions we gathered from Stack Overflow.

A. *Category of Question.*

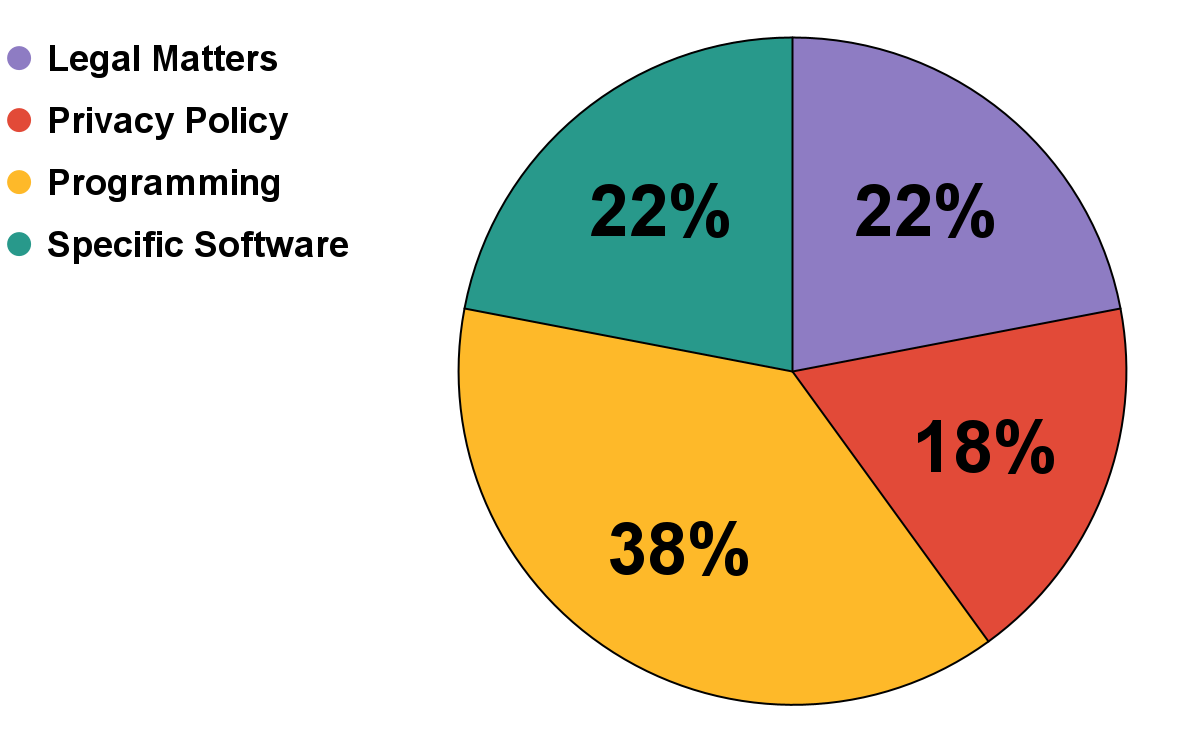
We found that each of the questions we gathered fell into four general categories: legal matters, privacy policies, programming, and specific software.

“Legal Matters”, was defined as having to do with compliance with regulations. Notably, most of the “legal matter” questions concerned how to comply with privacy legislation, especially GDPR; which alone accounts for 17 of the 22 legal questions.

“Privacy Policy”, related to how many people had questions about specifically how to create privacy policies that would comply with the rules set for by the Apple App Store and/or Google Play Store. Others had questions about the creation of privacy policies based on the specifics of the software they were writing.

“Programming” questions are about the actual writing of the software. Given that Stack Overflow is a site that is largely used to find answers to questions about coding, it’s no surprise that this was the largest of the four categories. Some of the programming languages mentioned in the questions include Java, C, C++, Python, R, HTTP, and Javascript. A large portion of these questions asked for advice on how to tackle a programming challenge in general and did not mention or give any example of a specific language.

“Specific Software” referred to how some people had questions that did pertain to specific software. Unlike the “Programming” questions, these questions are reliant upon something specific in order to be answered effectively. As an example, some were technical questions about APIs; while others were questions about the security, functionality or privacy of a piece of software.



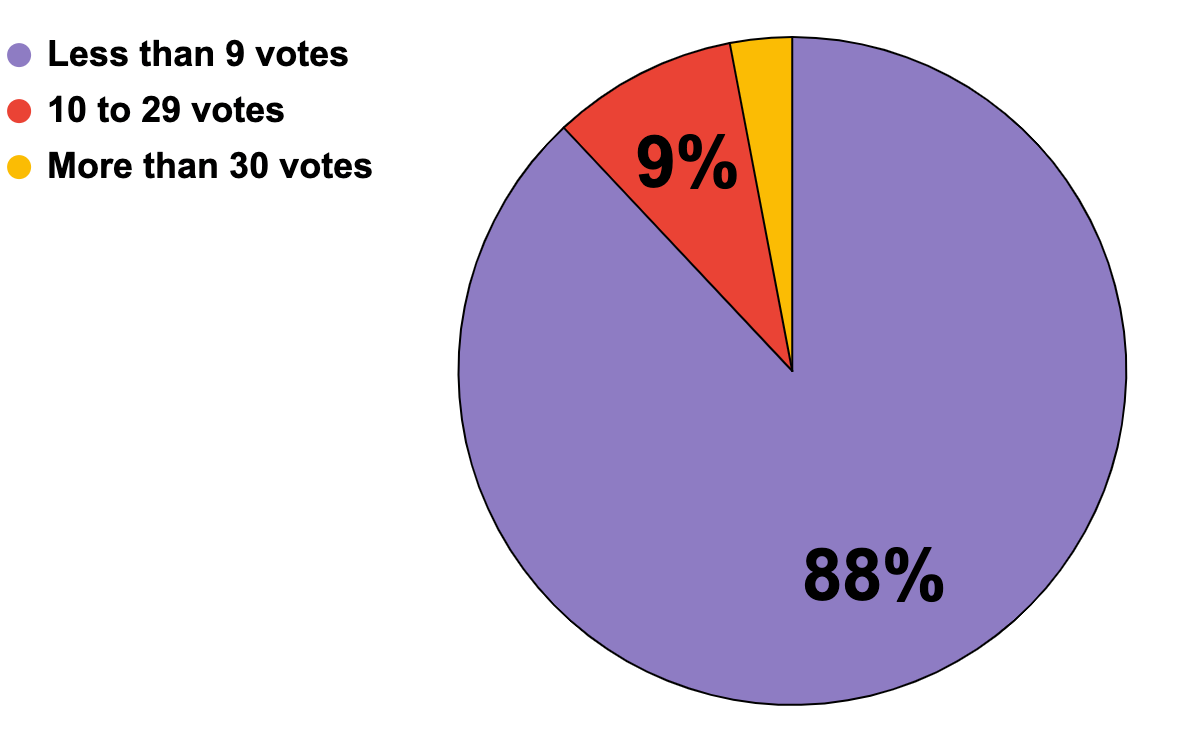
B. *Question Time-Frame.*

All of the questions that we collected for our study were posted no more than five years ago, so naturally none of the questions are from prior to the introduction of GDPR. It’s worth noting that only 18% of questions originate from prior to the enactment of GDPR. This means that rather than proactively preparing for these new regulations, it may be the case that developers only took action once the new regulations were fully implemented.

| **Question Time-Frame (Relative to Privacy Legislation)** | |
| --- | --- |
| **%** | *Privacy Legislation.* |
| **0%** | Questions prior to the introduction of the GDPR (2014 - 03 - 12). |
| **18%** | Questions were prior to the enactment of GDPR. (2018 - 05 - 25) |
| **13%** | Questions prior to the introduction of the CCPA. (2018 - 01 - 03) |
| **58%** | Questions were prior to the enactment of CCPA (2020 - 01 - 01) |

C. *Question scoring.*

Unfortunately, the vast majority of the questions had extremely low scores, which indicates that they did not have very much user-engagement. This may have impacted the quality and quantity of the answers. Only 3% of all questions had a score of greater than 30.



D. *Question views.*

As with the question scoring, it’s unfortunate such a significant number of questions on Stack Overflow on this topic have received such few views. It’s plausible that this level of viewership could have hampered the ability of the Stack Overflow community to effectively verify the answers that were given.

| **Relative Number of Views Received** | |
| --- | --- |
| **% of Questions** | *Number of views received per question.* |
| **61%** | Fewer than 1,000 views |
| **22%** | Between 1,0001 and 3,000 views |
| **17%** | More than 3,000 views |

E. *Question views by keyword.*

The questions we collected received 327,530 views in totality and the table below shows a breakdown of how those views were distributed based on their keyword. It was notable that the categories of “privacy” and “security” are considered common words which contain numerous sub categories which inflated their viewership more than the more specific categories of “COPPA” or “GDPR.

| **Percentage of question-views received (by keyword)** | |
| --- | --- |
| ***% of All Views*** *(num views)* | *Keyword(s)* |
| **2.6%** (8,547) | CCPA\* |
| **0.4%** (1,285) | CalOPPA |
| **0.4%** (1,235) | Compliant |
| **1.9%** (6,370) | COPPA |
| **12.1%** (39,561) | GDPR |
| **30.4%** (99,522) | Privacy\* |
| **52.2%** (171,008) | Security\* |

*\* The following keywords are composed of multiple closely-related words: CCPA (“California Consumer Privacy Act”, “California privacy”, “CCPA”), Privacy (“privacy”, “privacy act”, “privacy policy”), and Security (“safety”, “security”, “security policy”).*

F. *Question scoring by keyword.*

The questions we collected received 717 votes in totality and the table below shows a breakdown of how those votes were distributed based on their keyword(s). These scores were impacted heavily by user involvement, with higher view questions also containing a high score due to user satisfaction.

| **Percent of question-votes received (by keyword)** | |
| --- | --- |
| ***% of All Votes*** *(num votes)* | *Keyword(s)* |
| **4.9%** (35) | CCPA\* |
| **0.3%** (2) | CalOPPA |
| **0.7%** (5) | Compliant |
| **6.4%** (46) | COPPA |
| **14.5%** (104) | GDPR |
| **12.8%** (92) | Privacy\* |
| **60.4%** (433) | Security\* |

*\* The following keywords are composed of multiple closely-related words: CCPA (“California Consumer Privacy Act”, “California privacy”, “CCPA”), Privacy (“privacy”, “privacy act”, “privacy policy”), and Security (“safety”, “security”, “security policy”).*

G. *Question scoring versus number of views by keyword.*

There are several question keywords that performed significantly better than the others, most notably questions pertaining to CCPA, GDPR, and security. While “compliant” questions received a very high vote-to-view ratio, because our survey only collected two such questions, that number isn’t nearly as compelling. We were disappointed to see that even though “privacy” questions represent nearly one third of the questions we compiled, they received very few votes when compared to the number of views that they received.

Privacy may have received few votes for several reasons; it may be the case that users who visited those questions were less likely to have Stack Overflow accounts, and so would be unable to vote. It may also be the case that because “privacy” is such a broad topic, users may not have found what they were looking for, whereas a question about a specific regulation could be more helpful (i.e. GDPR, CCPA).

| **Number of question-votes received per 1000 views  (by keyword)** | |
| --- | --- |
| ***Num question-votes per 1000 views*** *(num questions)* | *Keyword(s)* |
| **4.10** (8) | CCPA\* |
| **1.56** (1) | CalOPPA |
| **4.04** (2) | Compliant |
| **2.79** (3) | COPPA |
| **2.63** (25) | GDPR |
| **0.92** (28) | Privacy\* |
| **2.53** (29) | Security\* |

*\* The following keywords are composed of multiple closely-related words: CCPA (“California Consumer Privacy Act”, “California privacy”, “CCPA”), Privacy (“privacy”, “privacy act”, “privacy policy”), and Security (“safety”, “security”, “security policy”).*

VI. Conclusions and Future Work.

In this paper, our team analyzed 100 questions and answers related to privacy and security on Stack Overflow- deriving common questions from users which related to legal advice, privacy policies, programming and the use of specific software and libraries. We discovered that many of these questions had wide differences in engagement, with most of the high engagement posts being related to programming and privacy policy generation. We also discovered that many of the questions and answers to these privacy policy questions bore similarities in structure and needs. Our team used these results to propose changes to Stack Overflow's infrastructure- such as a separate catalog for common questions which could be referenced by developers and more robust labeling procedures for questions related to the topics of security and privacy. In the future, this study could be used by other Stack Overflow technicians, or private researchers, as a backdrop and guideline for how to uncover and understand the concerns or problems facing developers on the Stack Overflow platform-or any other type of question board. In addition, our findings and conclusion could be taken as advice to future website creators and privacy boards as a means to further developer interests and understanding. This study could also serve as a backdrop for other future inquiries, such as “What other trusted resources are available to developers?”, “How do developers interact with digital-based legal advice on Stack Overflow?”, “What software is most commonly represented within questions related to security or privacy?”, and “How are privacy policies outside of Europe and America being represented on the website by developers?”

References.

[1] J. Atwood, J. Spolsky, P, Chandresekar, “Stack Exchange”, https://StackOverflow.com/ (accessed Sep. 17, 2022)

[2] M. Tahaei, K. Vaniea and A. Rashid, "Embedding Privacy Into Design Through Software Developers: Challenges & Solutions", in IEEE Security & Privacy, doi: 10.1109/MSEC.2022.3204364.

[3] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek and C. Stransky, "You Get Where You're Looking for: The Impact of Information Sources on Code Security," IEEE Security and Privacy (SP), doi: 10.1109/SP.2016.25.

[4] M. Tahaei, J. Bernd, A Rashid, “Privacy, Permissions, and the Health App Ecosystem: A Stack Overflow Exploration”, in EuroUSEC, doi: 10.1145/3549015.3555669

[5] M. Allamanis and C. Sutton, "Why, when, and what: Analyzing Stack Overflow questions by topic, type, and code," 10th Working Conference on Mining Software Repositories, doi: 10.1109/MSR.2013.6624004.

[6] N. Ferreyra, M. Vidoni, M. Heisel, R. Scandariato, “Cybersecurity Discussions in Stack Overflow: A Developer-Centered Analysis of Engagement and Self-Disclosure Behaviour”, in ARXIV Human-Computer Interaction, doi: 10.48550/arXiv.2207.01529

[7] G. L. Scoccia, P. Migliarini and M. Autili, "Challenges in Developing Desktop Web Apps: a Study of Stack Overflow and GitHub," IEEE/ACM 18th International Conference on Mining Software Repositories, doi: 10.1109/MSR52588.2021.00039.

[8] M. Tahaei, A. Frik, K. Vaenia, “Deciding on Personalized Ads: Nudging Developers About User Privacy,” in USENIX Seventeenth Symposium on Usable Privacy and Security, doi: 10.7488/ds/3045.

[9] T. Li, E. Louie, L. Dabbish, J. Hong, “How Developers Talk About Personal Data and What It Means for User Privacy: A Case Study of a Developer Forum on Reddit,” in ACM Journal on Human-Computer Interaction Vol.4, doi:10.1145/3432919

[10] K. Kubicek, A. Nielsen, D. Basin, “Query Tool for Analyzing Privacy Changes of GitHub-Hosted OSS Projects,” in ETH institute of Zurich, unpublished

[11] S. Mondal, M. Rahman, C. Roy. K. Shneider, “The Reproducibility of Programming-Related Issues in Stack Overflow Questions,” in ARXIV Software Engineering, doi: 10.48550/arXiv.2111.12204

[12] M. Tahaei, T. Li, K .Vaniea, “Understanding Privacy-Related Advice on Stack Overflow,” in SCIENDO Proceedings on Privacy Enhancing Technologies, doi: 10.2478/popets-2022-0032

[13] M. Tahaei, K. Vaniea, N. Saphra, “Understanding Privacy-Related Questions on Stack Overflow,” in ACM Conference CHI 20, doi: 10.1145/3313831.3376768

[14] CBR Staff Writer, “Hackers Gain “Production Access” to Stack Overflow,” in TechMonitor, https://techmonitor.ai/technology/cybersecurity/stack-overflow-hacked (accessed Sep. 18, 2022)

[15] R. Daws, “ Stack Overflow enhances ‘Content Health’ to end outdated knowledge,” in Techforge, https://www.developer-tech.com/news/2022/mar/04/stack-overflow-enhances-content-health-end-outdated-knowledge/ (accessed Sep. 17. 2022)

[16] Cyware Publishing, “Stack Overflow breach exposed user data”, https://cyware.com/news/stack-overflow-breach-exposed-user-data-44e05572 (accessed Sep 18. 2022)

x1

[17] European Union, “The EU General Data Protection Regulation (GDPR),” http://www.eugdpr.org, 2022 (accessed November 1, 2022)

[18] Government of California, “California Consumer Privacy Act (CCPA),” https://oag.ca.gov/, 2022 (accessed November 1, 2022)

**Contributions.**

| **Individual** | **Description of Work** | **Percent** |
| --- | --- | --- |
| **Anthony** | Completed the references section after feedback from deliverable #1.  Wrote multiple summaries for the literature review section.  Edited all of the summaries for the literature review section.  Corrected all of the summaries for the literature review section in accordance with feedback from deliverable #1  Wrote and edited the first draft of the abstract.  Wrote and edited the first draft of the introduction.  Collected 80 of the 100 questions for analysis.  Sorted 33 of the questions/answers and summarized them in latter documentation.  Gathered statistical data about the questions for analysis.  Wrote section A within results.  Wrote section B within results.  Wrote section B within discussion.  Corrected the abstract after feedback from deliverable #1.  Corrected formatting and addressed feedback from deliverable #2 across the entire document.  Wrote and edited conclusions and future work”.  Revised and edited section A within Results.  Revised and edited section B within Results.  Revised and edited section C within Results.  Revised and edited section D within Results.  Revised and edited section E within Results.  Revised and edited section F within Results.  Revised and edited section G within Results.  Created the Github for the project and uploaded all available files onto its space.  Formatted the Excel sheets for the scrubbed questions. | 33% |
| **Elijah** | Wrote majority of literature review summaries.  Proofread/Edited Abstract and Introduction section  Sorted/Summarized 33 of the questions  Wrote Section III  Revised Introduction  Edited Section V  Revised Section II  Edited Section IV | 34% |
| **Christian** | Assisted with the literature review sections.  Revised and proofread the Introduction section.  Edited and proofread the Abstract.  Assisted in formatting of the overall document.  Proposed metrics for data collection from Stack Overflow.  Sorted/Summarized 33 of the questions.  Wrote section IV.  Wrote section A within Results.  Created the chart for section A within Results.  Contributed to section B within Results.  Created the table for section B within Results.  Wrote section C within Results.  Created the chart for section C within Results.  Wrote section D within Results.  Created the table for section D within Results.  Wrote section E within Results.  Created the table for section E within Results.  Wrote section F within Results.  Created the table for section F within Results.  Wrote section G within Results.  Created the chart for section G within Results. | 33% |