Anonymity in Developer Communities: Insights from Developer Perceptions and Stack Overflow Profiles

Elim Yoseph Lemango

A Thesis Submitted to the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements for the Degree of
Master of Science

Graduate Program in Electrical Engineering and Computer Science

York University
Toronto, Ontario

September, 2024

Abstract

This thesis consists of two studies: an interview study with 34 early-career developers and a mining study analyzing 130,000 developer profiles. The interview study examines developers' definitions of anonymity, their preferences for anonymity, and their engagement with privacy policies. It also explores whether presenting privacy policies using contextual integrity principles improves understanding of the privacy policies. The developers from the interview study defined anonymity as withholding identifiable information like name, location, and professional background. The mining study investigates how much information developers share across platforms and the ease of retrieving their professional profiles. Our findings show that using Stack Overflow location and screen name in LinkedIn searches narrows down profiles, but cross-linking Stack Overflow profile data information with GitHub or Twitter adds noise. This research provides valuable insights into how developers define anonymity, and how that affects their behaviour when using social coding platforms.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to God, whose grace and guidance have been my source of strength throughout this journey.

I am very grateful to my supervisor, Dr. Maleknaz Nayebi, for her unwavering support, insightful guidance, and constant encouragement. Your expertise and patience have been instrumental in shaping this thesis, and I am truly fortunate to have had the opportunity to learn from you.

I would also like to extend my sincere thanks to a member of my supervisory committee, Dr. Yan Shvartzshnaider for his valuable feedback and suggestions that have greatly enriched this work. Your feedback has been crucial in helping me navigate the complexities of this research.

To my family, I am deeply thankful for your love, sacrifices, and endless encouragement. Your belief in me has been a constant source of motivation, and your support has given me the strength to overcome the many challenges of this academic journey.

I am also grateful to my friends for their understanding, companionship, and encouragement. Thank you for being there through the highs and lows, and for always believing in me.

Finally, I would like to thank my lab colleagues for creating a collaborative and supportive environment. The discussions, feedback, and shared experiences have greatly contributed to the completion of this thesis. Your friendship has made this journey not only bearable but truly enjoyable.

Table of Contents

Abstract . . . . . **ii**

Acknowledgments . . . . . **iii**

Table of Contents . . . . . **iv**

List of Tables . . . . . **vii**

List of Figures . . . . . **viii**

1 Introduction **1**

1.1 Current Knowledge on Developer Anonymity 2

1.2 Anonymity and Developer's Profile on Social Coding Platforms 3

1.3 Privacy Policies and Contextual Integrity on Social Coding Platforms . . . 5

1.4 Thesis Overview 8

2 Anonymity and Privacy Policies: Software Developer Perceptions and Engagement on Social Coding Platforms **9**

2.1 Introduction 9

2.2 Related Works 11

2.2.1 Anonymity on social media platforms 11

2.2.2 Anonymity on social coding platforms 11

2.2.3 Anonymity and Privacy policy 13

2.3 Research Methodology 14

2.3.1 Interview Guide 14

2.3.2 Interview Protocol 15

2.4 Interview Study Results 18

2.4.1 Interview study participants 18

2.4.2 Definition of anonymity by developers (RQ1) 19
2.4.3 Self-perceived anonymity status (RQ2) 24
2.4.4 Developers' engagement with privacy policies (RQ3) 32
2.4.5 Impact of context on privacy policy understanding (RQ4) 37
2.5 Discussions and Implications 41
2.5.1 Impact of intersectional diversity factors on the urge for anonymity 41
2.5.2 Does Gender Identity have an Impact on Anonymity? 42
2.5.3 Perception of Anonymity Compared to Reality 43
2.5.4 Triangulating Developer Activity and Profile Completeness on Stack Overflow 44
2.5.5 Triangulating Developer Activity and Profile Completeness on GitHub 45
2.6 Threats to Validity 46
2.6.1 Conclusion Validity 47
2.6.2 Construct Validity 47
2.6.3 Internal Validity 48
2.6.4 External Validity 48
2.7 Conclusion and Future Work 48

3 Enigma of Anonymity on Social Coding Platforms 50
3.1 Introduction 50
3.2 Related Works 52
3.3 Empirical Data and Method 54
3.3.1 Completeness of developers' profile (RQ1) 54
3.3.2 Identifying LinkedIn profiles (RQ2) 55
3.4 Empirical Results 56
3.4.1 Completeness of developers' profile (RQ1) 56
3.4.2 Identifying LinkedIn profiles (RQ2) 57
3.5 Discussions and Emerging Research 60
3.5.1 Is developers' anonymity intentional or not? 60
3.5.2 Is anonymity beneficial for developers' careers or not? 60
3.5.3 Why do developers feel the need to be anonymous? 62

3.5.4 Does consistency of the information developers share play a part in anonymity? . . . 63
3.6 Threats to Validity . . . 69
3.7 Conclusion . . . 70
3.8 Future Work . . . 70

4 Conclusion . . . 71
4.1 Future Work . . . 72

Bibliography . . . 74

Appendix A: Interview Study Questions . . . 89

Appendix B: Stack Overflow Privacy Policy . . . 96

Appendix C: GitHub General Privacy Statement . . . 108

List of Tables

2.1 Demographics results of surveying 34 early career software developers. . 18
2.2 Questions and interview survey results of 34 developers with the objective RQ1:"*Definition of anonymity by developers*" 23
2.3 Questions and interview survey results of 34 developers with the objective of "*RQ2: Self-perceived anonymity status*" 25
2.4 Questions and results for 34 developers interview RQ3 objective "*Developer engagement with privacy policies*" 27
2.5 Important statements for the developers from Stack Overflow's privacy policy (Q23). 31
2.6 Important statements for the developers from GitHub's privacy policy (Q25). 32
2.7 Questions and results for 34 developers interview RQ4 objective "*Impact of context on privacy policy understanding*" 36
2.8 Stack Overflow activity metric with the definition of the metric. 43
2.9 Percentage of anonymous accounts and multiple accounts of the developers with gender 44
2.10 Perceived vs. Actual Anonymity on Social Coding Platforms 44
2.11 GitHub activity metric with the definition of the metrics. 45

3.1 Search strings to retrieve LinkedIn profiles in **RQ2**, using Stack Overflow (), GitHub (), and Twitter () data. 55
3.2 Identified users on LinkedIn () with Stack Overflow (), GitHub (), Twitter() and aggregated attributes 58
3.3 Results of surveying 117 software developers to determine if they are identifiable or anonymous on Stack Overflow 61
3.4 Consistency of Text & Image Attributes on , , ... 68
3.5 Comparison of Stack Overflow and LinkedIn users Image Attributes , 69

List of Figures

1.1 Profile dashboard for “Elmira” (Added by consent) 4
1.2 Profile Dashboard for “Sharuka Thirimanne” (Added by consent) 6
1.3 Profile Dashboard for “Nani” (Added by consent) 7

2.1 Methodology of our interview study. 14
2.2 Attributes important for developers to protect through anonymity (Q6). . 19
2.3 Platforms developers have anonymous accounts (according to their own definition): (a) social media, (b) social coding and professional platforms (Q11). 21
2.4 Importance of anonymity for the developers on Stack Overflow and GitHub (Q13). 28
2.5 Platforms developers use real profile pictures on: (a) social media, (b) social coding and professional platforms (Q14). 30
2.6 Developers have multiple accounts on: (a) Social media platforms, (b) Social coding and professional platforms (Q17). 33
2.7 Likeliness of reading Stack Overflow’ and GitHub’s privacy policy (Q22) and (Q25). 34
2.8 Likeliness of reading Stack Overflow’s and GitHub’s privacy policy before and after applying Contextual Integrity on the policy (Q22), (Q25), (Q30), and (Q31). 35
2.9 Heat map of the Stack Overflow privacy policy ranking provided by the developers for the documents that were restructured using Contextual Integrity 37
2.10 Heat map of the GitHub privacy policy ranking provided by the developers for the documents that were restructured using Contextual Integrity . 38
2.11 Stack Overflow activity metric with Stack Overflow attribute availability. . 45

2.12 GitHub activity metric correlation with GitHub attribute availability. . . . 47

3.1 Availability of “optional" attributes on active sampled Stack Overflow (a), GitHub (b), and Twitter (c) profiles. 55

3.2 Findablity of unique profiles on LinkedIn 59

Chapter 1

Introduction

Developers do not always reveal their true identities on public online social coding platforms. However, platforms like social coding platforms are increasingly adopting methods for recognition and engagement [64, 123]. From a career development perspective, linking recognition to their identity can benefit developers, as recruiters and potential employers often review these profiles [124]. Social coding platforms help developers build an online portfolio by engaging in open-source projects and knowledge crowdsourcing, enhancing their career reputation [60].

Social coding platforms, such as Stack Overflow[1] and GitHub[2], have revolutionized software development by enabling crowd sourcing knowledge, collaborative coding, issue tracking, and version control [110]. Stack Overflow is a question and answer website developers use to ask and answer programming-related questions, which hosts 39 million users [9]. GitHub is one of the most popular open-source collaborative coding sites, which hosts about a million users [39]. These platforms provide a space where developers from around the world can contribute to projects, share their expertise, and build professional reputations. They facilitate open-source development, where transparency and community contributions are highly valued. The collaborative nature of these platforms encourages a sense of community and collective problem-solving, driving innovation and knowledge sharing.

Studies show that extrinsic motivations, especially related to career prospects, directly motivate users to contribute, and gamified incentive mechanisms help achieve this goal

[1] `https://stackoverflow.com`
[2] `https://github.com`

[63, 17, 115]. In recent years, social coding platforms have revolutionized how developers collaborate, share knowledge, and contribute to open-source projects [17, 115]. Platforms like Stack Overflow and GitHub facilitate software co-development, knowledge sharing, and professional collaboration. These platforms enable developers to build an online portfolio, engage in open-source projects, and crowd-source knowledge, enhancing their career reputation [60].

1.1 Current Knowledge on Developer Anonymity

There is no one definition of anonymity for software developers [37]. In the broader literature, anonymity has various definitions that are used throughout different domains. In its essence, anonymity in the online world is explained as the notion of being seen by others but not being identifiable [37]. This means that even though people see the activities one does online, they won't be able to accurately link that to a real-world person. To understand anonymity, we would need to take into account both the technical and social dimensions of the context at hand [92]. In the technical sense, this refers to the ability of a person to interact in online spaces without their real identity being exposed or linked to their activities. From a social perspective, anonymity affects the dynamics of communication, interaction, and collaboration in online communities. Hence, we can't conceptualize it in absolute terms but as a fluid concept that can characterize social online interactions for software developers [92].

Also, the use of real identity and its impact on developers' careers is controversial among researchers and practitioners. One of the reasons there is tension between researchers and practitioners is that absolute anonymity can result in developers feeling less accountable for their actions, it can lead to the spread of misinformation and discourage constructive and respectful interactions [87]. On the other hand, absolute identification on a platform requires its users to reveal their entire identity, including name, location, and other personal details [87]. This can help to foster community trust and encourage more accurate and constructive interactions. However, if a developer's identity is revealed in a negative context, it may expose them to potential risks such as discrimination or harassment [37]. Pseudonymous identification, for example, allows people to

use a nickname-based screen name. This can provide a false sense of privacy while still allowing for some trust, accountability, and transparency within an online community.

Ultimately, the appropriate level of anonymity depends on the context and the needs of the individuals involved [82]. Anonymity is a personal choice, shaped either by individual preference or imposed by a platform's organizational policies. It is highly context dependent, as individuals' diverse backgrounds, experiences, and identities influence their expectations and needs for privacy. For example, a software developer contributing to an open-source project may value anonymity differently than a developer seeking recognition for their work or career advancement. The individual's identity such as their gender, race, cultural background, or professional standing can affect how they perceive anonymity and what level of protection they seek in online environments [117].

Moreover, in online platforms, particularly in technical communities like social coding platforms, the decision to remain anonymous or partially identifiable often relates to concerns around safety, inclusivity, or bias [117]. Marginalized individuals, for instance, may seek higher levels of anonymity to avoid harassment, discrimination, or biases based on their real-world identity. On the other hand others may choose partial anonymity or pseudonymity to balance personal safety with professional recognition or credibility. It is important for individuals and communities to carefully consider the benefits and drawbacks of different levels of anonymity to strike a balance that promotes privacy, trust, and constructive interactions.

1.2 Anonymity and Developer's Profile on Social Coding Platforms

In this section, we present an initial analysis of the Stack Overflow profiles where developers can present themselves using the user profile page [61]. Additionally, we present motivating examples of how users can either enhance or diminish their anonymity based on their online behavior. By completing their profiles or consistently sharing personal information, users may unintentionally reveal identifiable patterns that undermine their anonymity. For instance, adding details like real names, photos, or professional affiliations can make it easier for others to link their online activity to their real-world identity,

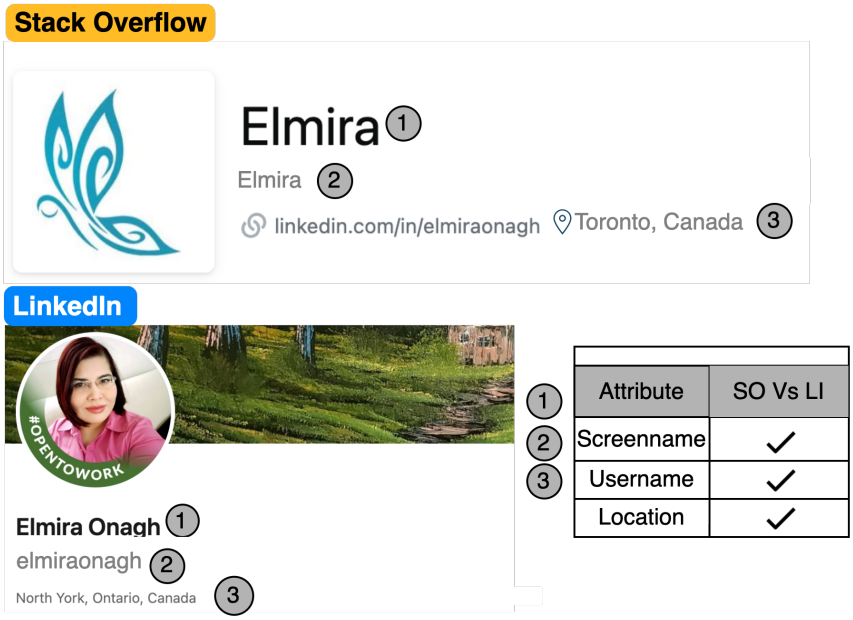


FIGURE 1.1: Profile dashboard for "Elmira" (Added by consent)

increasing linkability. On the other hand, users who carefully manage their online presence limiting what they share and using pseudonyms—can maintain stronger anonymity. Thus, anonymity and linkability are fluid and can shift depending on how much information a user chooses to disclose over time. Stack Overflow allows developers to showcase comprehensive details in their user profiles, alongside the achievements garnered through their platform activities, such as badges and reputation points [61]. While the screen name (display name) and username are the only mandatory information on Stack Overflow, developers can enhance their profiles by adding further information such as location, avatar (image), and even further linking GitHub, Twitter, and personal websites. However, it's noteworthy that developers don't consistently provide identical details across these linked accounts nor do they consistently link these accounts.

For example, the developer with the screen name "Elmira" has not provided links to their GitHub or Twitter profiles on Stack Overflow which can contribute to her being anonymous. Conversely, they have included a link to their LinkedIn profile on Stack Overflow. On Stack Overflow, their displayed screen name is "Elmira," while "Elmira Onagh" is shown on LinkedIn (See Figure 1.1). In contrast, "Sharuka Thirimanne" exemplifies consistency in their online identity which acts as a barrier to his anonymity. Their chosen screen name, "Sharuka Thirimanne" remains consistent across Stack Overflow, GitHub, and Twitter. Similarly, variations of the username "sharukathirimanne"

are consistently used on both GitHub and Twitter. Concerning location, Sharuka specifies "Toronto, Canada" on Stack Overflow, "Toronto, Canada" on GitHub, and "Vaughan, Ontario" on Twitter. Although the locations are not an exact match, there is a notable degree of coherence across his diverse online accounts (See Figure 1.2). A straightforward search of their screen name on LinkedIn yields two results at the time of writing this paper. However, when combining the screen name with the location, such as "Toronto", this search is narrowed down to only one profile which is consistent with their Stack Overflow, GitHub and Twitter.

The developer using the screen name "Nani" on Stack Overflow has provided links to their GitHub and Twitter accounts. However, while they go by "Nani" on Stack Overflow, their screen names on GitHub and Twitter are "Elim Lemango" and "ElimYLemango," respectively. Additionally, the username "nani" is used on Stack Overflow, whereas "elim177" and "elimylemango" are consistently used on GitHub and Twitter. Location information is consistently shared on Stack Overflow and GitHub but is inconsistent on Twitter (see Figure 1.3). A LinkedIn search for this screen name returns 74 profiles, only one of which uses the screen name "Elim Lemango," matching the one provided on their GitHub and Twitter accounts.

1.3 Privacy Policies and Contextual Integrity on Social Coding Platforms

Stack Overflow and GitHub facilitate collaboration among developers. Given the collaborative nature of these platforms, privacy policies play a crucial role in protecting user data, including personal information and code contributions. Given the amount of publicly shared code, discussions, and potentially identifiable profiles, privacy policies play a crucial role in protecting user data. These policies must not only protect personal information such as names, email addresses, and locations but also address the ownership and security of code contributions, which are often closely linked to a developer's professional identity. However, research indicates that privacy policies are often complex and challenging to understand, leading to widespread non-compliance or misunderstanding among users [106]. The contextual integrity framework, proposed by Helen Nissenbaum, provides a way to assess privacy practices by examining how information flows conform

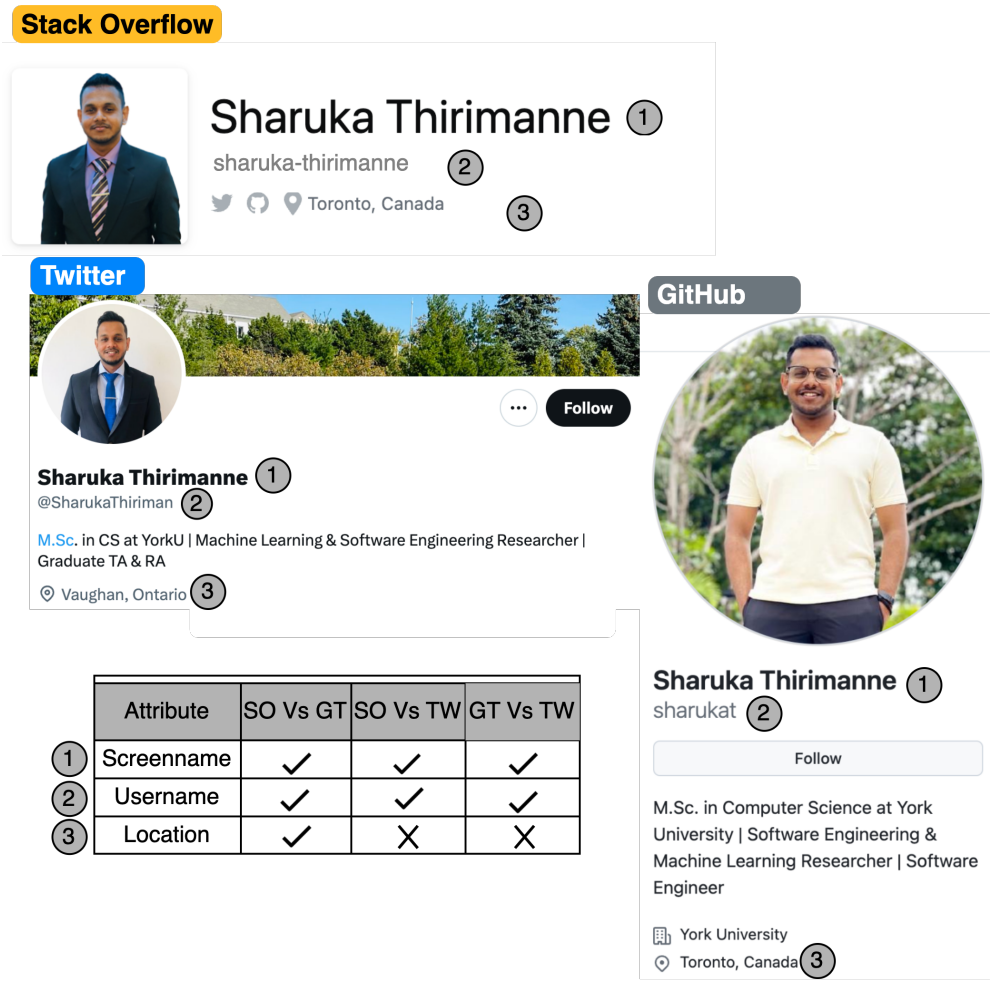
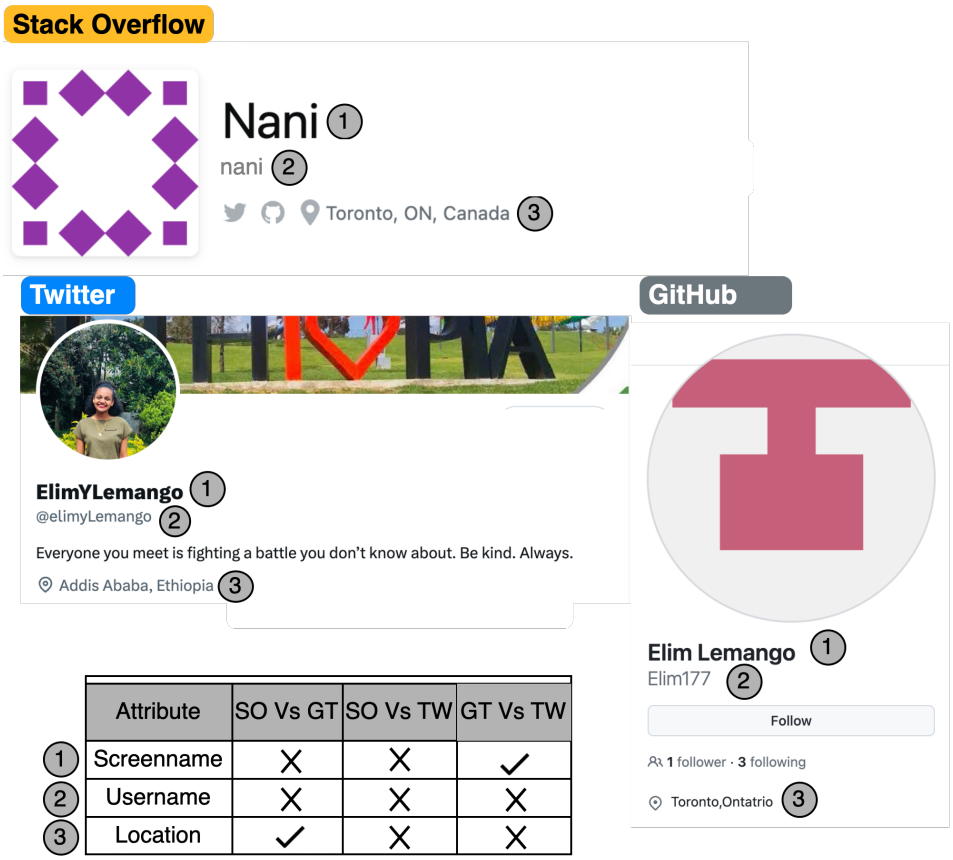


FIGURE 1.2: Profile Dashboard for "Sharuka Thirimanne" (Added by consent)

to social norms within specific contexts. Applying contextual integrity to privacy policies on social coding platforms could enhance developers' understanding and engagement by aligning these policies with the expectations and norms of the developer community. This approach could be more effective than the existing formats (current formats), which tend to focus on legal compliance rather than user comprehension [130].

Recent studies have also highlighted the impact of the General Data Protection Regulation (GDPR) on privacy practices across various digital platforms, including social coding environments. The GDPR emphasizes transparency, data minimization, and user consent, compelling platforms to update their privacy policies to meet these standards [24]. Despite these advancements, developers often find it challenging to navigate and apply these policies, particularly when extracting specific privacy options like opt-out statements [8]. Thus, while the GDPR has driven significant improvements in privacy



	Attribute	SO Vs GT	SO Vs TW	GT Vs TW
1	Screenname	✗	✗	✓
2	Username	✗	✗	✗
3	Location	✓	✗	✗

FIGURE 1.3: Profile Dashboard for "Nani" (Added by consent)

protection, its implementation in developer-centered platforms requires further refinement to ensure that developers not only understand but also effectively engage with these policies.

In this thesis, we explored how developers define and perceive anonymity on social coding platforms, their engagement with privacy policies, and the impact of presenting privacy policies in different formats. To this end, we investigated the definition of anonymity from the perspective of early-career software developers.

We recognize the existence of the different ways developers might act on a platform based on the context of the platform [69]. This research is important in a time when online collaboration is essential and open-source community dynamics have a significant impact on the direction of technology [69]. We can obtain an understanding of the many motivations and behaviors that drive collaborative software development by understanding the specifics of developer anonymity.

1.4 Thesis Overview

This thesis is manuscript-based and consists of two papers. The first paper is based on an interview study to define anonymity from the perspective of 34 early-career developers. Surveys and interviews are established means for the exploration of new notions [79, 77, 80, 78, 94, 58, 57, 73]. This paper aimed to investigate early career developers' definition of anonymity and determine their engagement with privacy policies. We aimed to understand developers' self-perceived definition of anonymity (RQ1). We further explored the reasons why developers seek anonymity and the specific activities in which it is preferred (RQ2). Moreover, we aimed to determine developers' engagement with the privacy policies of Stack Overflow and GitHub (RQ3) by presenting them with "Personal Data" sections from these policies. Lastly, we investigated whether contextual integrity improved developers' understanding of privacy policies (RQ4) by annotating the policies and restructuring the policies.

The second paper is a mining study that was motivated by the interview study we performed to see how easily developers' professional profiles can be retrieved using their profile attributes from social coding platforms. Specifically, we determined the extent to which developers share their identifying information on coding platforms and social networks (RQ5). To achieve this, we sampled 130,000 developer profiles on Stack Overflow and analyzed their profiles across different platforms. We also aim to assess how easily developers' professional profiles on LinkedIn can be retrieved using the information shared on social coding platforms (RQ6), by developing a search and retrieval process based on the data from RQ5.

In Chapter 2, we present our first paper which is the interview study with the 34 early-career developers, exploring how they define and perceive anonymity on social coding platforms. This chapter also explores their engagement with privacy policies and the influence of contextual factors on their understanding of these policies.

In Chapter 3, we present the second paper which is the mining study. We discuss the extent to which developers share identifying information on social coding platforms and assess how effectively their social coding platform attributes can be used to retrieve their professional profiles. In Chapter 4, we present the final conclusions of this thesis and discuss potential areas for future research.

Chapter 2

Anonymity and Privacy Policies: Software Developer Perceptions and Engagement on Social Coding Platforms

2.1 Introduction

Software developers are increasingly utilizing social coding platforms for collaboration, knowledge sharing, and career development [60]. According to global data and business intelligence platform *Statista*, the software professional population has grown by over 20% in the past five years and is expected to reach 28.7 million in 2024[1]. These platforms enable developers to engage in global communities, where they contribute code, seek help, and build their reputations. Anonymity among software developers is a complex issue, as it affects interactions, collaboration, and the perception of one's contributions [37]. For many, anonymity serves as a protective mechanism, allowing them to engage without fear of judgment based on their background, experience level, or personal characteristics. This is the norm for early-career developers or individuals from underrepresented groups in tech, who may face bias or prejudice when presenting themselves in these platforms. Anonymity allows them to focus on technical merit rather than being evaluated based on visible demographic markers such as gender, race, or nationality [34]. Hence, anonymity can create a more equitable space, where diverse voices contribute to the open-source ecosystem without the barriers of biased judgments.

[1] `https://www.statista.com/statistics/627312/worldwide-developer-population/`

The perspectives of early-career developers and experienced developers on anonymity and privacy may differ [98]. Early-career developers, especially those from underrepresented groups, may rely on anonymity to contribute more freely without fear of prejudice or discrimination[98, 26]. For these early-career developers, anonymity provides a level playing field where the quality of code and contributions is prioritized over visible identity factors. This sense of protection can encourage greater participation and help individuals build confidence in their technical abilities without the pressure of public judgement. Research shows that underrepresented developers often face higher barriers in open-source communities, making anonymity a key factor in reducing bias [34].

In contrast, experienced developers might exhibit a more cautious approach, valuing their established professional identity and being more aware of the potential privacy implications [26]. They might be more selective about the information they share and the tools they use, prioritizing control over their digital footprint. Early career software developers are essential to the software engineering community [128]. For this paper, we define early-career developers as current students who have less than five years of experience. Various research studies have explored early-career developers and their workplace behavior, examining their attitudes and the link between their personalities and work preferences [45, 2].

Motivated by the lack of research from the perspective of early-career developers and their views on anonymity, we conducted an interview survey to define anonymity from their perspective and understand how they perceive their own level of anonymity. In this context, we addressed four research questions:

RQ1- How do Software developers define anonymity on social coding platforms?

RQ2- How do Software developers perceive their own anonymity on social coding platforms?

RQ3- To what extent do developers engage with privacy policies on social coding platforms?

RQ4- How does presenting privacy policies in contextual integrity (CI) principles affect developers' understanding compared to traditional formats?

2.2 Related Works

In this section, we present a thorough review of existing literature to understand the definition of anonymity on social media platforms and social coding platforms. We also present the relationship of anonymity with privacy policies.

2.2.1 Anonymity on social media platforms

Since the age of the Web, anonymity is considered one of the most essential features of social media [68]. It creates a condition under which individuals can freely express their views without having to be concerned about public pressure or regulatory repression, which in turn promotes discussion of controversial issues [49, 68]. Social media anonymity may also enhance individuals' psychological well-being by increasing their sense of control over privacy and personal boundaries [21]. However, social media anonymity reduces the cost of inappropriate behavior and makes it difficult to prohibit cyber-aggression [27, 43]. There has been previous research that has tried to investigate the complex interplay between anonymity and identity curation on social media platforms [109, 113]. Identity curation has been defined as the way people present themselves online in terms of their profile attributes such as their avatars.

As social media platforms go people are on there as social beings who act according to the platform they are on. A key part to this sociability is our ability to change and adapt our behavior in response to both internal (e.g., emotion, expectations) and external (e.g., audience, environment) factors [42]. This social flexibility extends to our behavior across contexts—we would not expect a person to behave identically as they transit from work to home or from home to a night out with friends. This type of context-based behavioral change is normal and expected by those around us [33, 41]. This may also be seen in an online context, where a user will behave differently as they move from one online platform to another.

2.2.2 Anonymity on social coding platforms

Social coding platforms and the interaction among developers have been the subject of many studies to provide actionable insight [74, 76, 91, 72]. According to Marx [66], being anonymous means a person cannot be identified according to any of seven dimensions

of identity knowledge, that is, the person's legal name, location, pseudonyms that can be linked to the person's legal name or location, pseudonyms that cannot be linked to specific identity information but that provide other clues to identity, revealing patterns of behavior, membership in a social group, or information, items, or skills that indicate personal characteristics. In terms of prior research that has been done there are two different views of anonymity on social coding platforms [30].

Anonymity and identification are important attributes that exist together on these social collaborative platforms. Anonymity provides users with privacy, allowing them to ask questions or share information without fear of consequences, while identification fosters trust within the community and promotes more precise and constructive interactions [82]. Open collaboration platforms that permit anonymity can offer a safer environment for individuals who may encounter discrimination or backlash due to stigmatized or marginalized identities [12, 48, 31]. While, developers who are not anonymous will focus more on the validity of the contributed knowledge, and produce more helpful knowledge, leading to a higher collaboration efficiency [5, 84]. Despite these measures, it is crucial to acknowledge that not all developers may feel comfortable revealing their identity, emphasizing the need to define and understand the spectrum of anonymity on platforms social coding platforms. Anonymous participants lack incentives for their contributions and have less responsibility to finish tasks, this can lead to deception in open collaboration and delays in task completion [89, 23].

Research shows that online collaboration on social coding platforms yields more tangible outcomes when it includes participants from varying backgrounds [6, 111]. This diversity brings a wide range of perspectives and experiences, enriching the collaborative process. Anonymity on social coding platforms plays a crucial role in fostering this diversity. By allowing contributors to remain anonymous, these platforms encourage individuals to share their opinions and ideas without fear of judgment or bias. This leads to more equal and unbiased participation, as developers are evaluated based on the merit of their contributions rather than personal attributes such as their identity, status, or background [59, 100]. As a result, the collaborative environment becomes more inclusive, enabling a richer exchange of ideas and more innovative solutions.

To address these tensions, social coding platforms such as Stack Overflow and GitHub have implemented an identification system based on user profiles [129, 62]. Developers

can create profiles with details about their education, work experience, and areas of expertise. This allows other users to assess an individual's credibility and expertise, which can help improve the quality and accuracy of the information shared on the platform. Furthermore, user profiles can aid in the development of community trust, as users are more likely to interact constructively when they have a reputation to uphold. Regardless of the advantages of identification, it is important to recognize that not all users will feel comfortable revealing their identity on Stack Overflow. Some people still prefer to remain anonymous for privacy or security reasons, and it is critical that the platform respects these concerns and allows for anonymous participation [105, 71]. To that end, in this paper, we will define anonymity from the perspective of early-career software developers and ask them about their perceived anonymity level on social coding platforms.

2.2.3 Anonymity and Privacy policy

A privacy policy is a document, which outlines how service providers process users' or customers' personal data [19]. Privacy regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) have been enacted in recent years, obliging software developers to take actions to protect users' privacy [10]. The existence of these laws does not negate the fact that privacy violations are still common in today's platforms and applications [35, 67]. Researchers have created frameworks to help synthesize guidelines and strategies for incorporating privacy values from the initial stages of design, such as Privacy by Design [15, 16] into the existing platforms and applications. Privacy by Design framework was originally proposed in 2009 [15], and was a reference for building privacy-friendly systems [13]. A lot of efforts have been made to translate "Privacy by Design principles" into concrete design guidelines, such as the privacy design strategies by Hoepman [44]. However difference between the increase of privacy framework research and the spread of privacy violations calls for systematic research into developers' awareness, the usefulness, and operationality of these frameworks [121].

There is not much known on the perspective of software developers on the importance of using privacy policies on their day to day tasks. Privacy policies are essential in order to inform users not only about personal data but also a user's ability to remain anonymous on platforms. For software developers, particularly on social coding

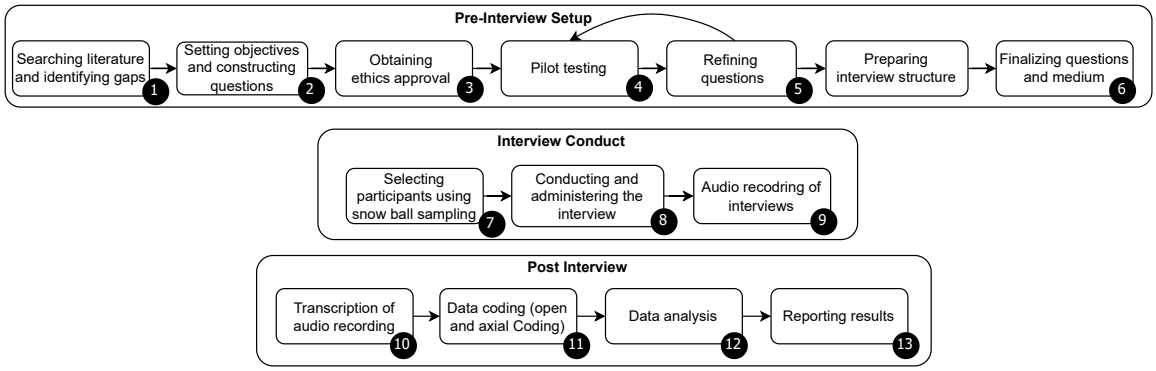


FIGURE 2.1: Methodology of our interview study.

platforms, privacy policies help define what personal information is collected, how it is used, and what measures are in place to protect their identity. Our research stands out by examining anonymity through a fresh lens of early-career developers, a user group not typically considered in this context. Furthermore, we focus on social coding platforms, a relevant online environment for this population. Our work also explores how developers engage with privacy policies. By clearly outlining the data collection practices, privacy policies allow users to make informed decisions when interacting on platforms. Finally, we investigate the potential of presenting privacy policies with the Contextual Integrity framework [82], a method that could improve user understanding specifically among software developers. By focusing on these under-explored areas, our research offers a unique contribution to the understanding of anonymity and privacy within the software development community.

2.3 Research Methodology

To determine the answers to our four research questions we carried out an interview survey which we elaborate in the following section. The main objective of the interview is to explore the meaning of anonymity for software developers particularly in relation to platforms' privacy policies.

2.3.1 Interview Guide

The interview includes four demographic questions to understand participants' professional roles, residence, experience, and gender. We collected their demographic information to understand its potential influence on anonymity preferences. The main part of the

interview consists of the questions that helped us address the objectives of our interview study:

Definition of anonymity by developers: We asked a series of six open-ended questions to determine the developers' self-perceived definition of anonymity. The purpose of these questions was to answer **RQ1**.

Self-perceived anonymity status: We asked a series of eleven questions aimed to uncover developers' self-perceived reasons for seeking anonymity and the specific activities in which it is preferred. The purpose of these questions was to answer **RQ2**.

Developers' engagement with privacy policies: We asked the developers a series of eight questions to understand their level of engagement with the privacy policy of Stack Overflow and GitHub to answer **RQ3**. We presented the participants with "*Personal Data*" section from the privacy policies of Stack Overflow and GitHub.

Impact of context on privacy policy understanding: The privacy policies were annotated by the first author using the definition of privacy in terms of contextual integrity. We asked the developers a series of four questions along with the restructured privacy policies to see if it improved their understanding of the privacy policy. The purpose of these questions was to answer **RQ4**.

2.3.2 Interview Protocol

In this section, we explore the pre-interview process, followed by the interview itself, and conclude with the post-interview steps for data analysis (See Figure 2.1).

Pre-Interview Setup

Before the interview, we conducted a thorough review of literature, set objectives, obtained ethics approval, performed pilot testing, refined the questions, and prepared the interview structure to finalize the questions and medium for the study.

❶ **Searching literature and identifying gaps**: We carried out a comprehensive review of relevant research papers to gain insights into ways to perform interview studies [88, 55, 56, 50, 51, 52, 75]. This review helped us to formulate the interview questions and identify existing interview protocols used in similar studies.

❷ **Setting objectives and constructing questions**: Based on the research questions and literature review conducted, we developed a set of semi-structured interview questions [70]. We conducted a semi-structured interview because they are frequently used as a data collection technique within the field of software engineering even though they involve high costs, since the quality of the collected data is higher [70]. The interview questions included the main areas to be explored while allowing flexibility for follow-up questions and in-depth exploration.

❸ **Obtaining ethics approval**: We also applied and received approval from the Ethics approval board at York University to carry out the interview study.

❹ **Pilot testing**: We conducted a pilot test with a small group of software developers in our close network (convenience sampling) to ensure clarity, flow, and effectiveness in capturing desired information from the interview survey [95]. The initial test was done with two developers who were graduate students. This pilot testing helped us refine the question's wording and guaranteed the interviewer (first author) covered the intended scope.

❺ **Refining questions**: Iteratively, we tested and refined the questions using two participants so that the questions were clear. We also timed the two interviews and estimated an average of 40 minutes needed for each interview.

❻ **Preparing interview structure**: The interview guide, which is the set of interview questions [70] served as a framework, ensuring information related to our initial research questions was gathered. We also wanted to allow for flexibility to adapt to individual responses and pursue emerging themes.

❼ **Finalizing questions and medium**: Based on pilot testing feedback, we sharpened the interview questions for improved clarity.

Interview Conduct

This includes the recruitment, atmosphere, type of interview, length of the interview as well as the method of capturing responses for the interview we conducted [36].

❽ **Selecting participants using snowball sampling**: We used a non-probabilistic sampling method known as Snowball sampling to recruit participants. This approach relies on initial contacts from early career software developers who then refer their peers and colleagues within their networks [36]. This technique is effective for reaching niche

populations like early-career developers who might not participate in open recruitment channels. We conducted interviews with 34 software developers.

❾ **Conducting and administering the interview**: Before starting the interview, we asked participants for their consent by informing them of the purpose of the interview, data collection procedures, confidentiality measures, and their right to withdraw their consent at any stage of the interview. We conducted in-person Semi-structured interviews by the first author when possible and online Zoom interviews otherwise.

❿ **Audio recording of interviews**: With the permission of participants, we audio-recorded the interviews with participant consent to ensure accurate capture of responses for later transcription and analysis. Each interview lasted approximately 43 minutes on average which was sufficient time for in-depth exploration without participant fatigue.

Post-Interview

After the interview we carried out a transcription of the audio recordings, and analyzed the texts by coding them.

⓫ **Transcription of audio recording**: Audio recordings were transcribed verbatim to create a text-based record of each interview. We used a transcription service called `Transcribe`[2] to transcribe the audio from the interviews. The interviewees were made aware that this was a necessary step before the start of the interview.

⓬ **Data coding**: We applied an iterative process of coding the interview transcripts by following the approach of Treude [107]. We used open coding and axial coding to carry out our qualitative analysis. Open coding in qualitative analysis allows loose and tentative coding to identify emerging concepts or themes. Axial coding is a technique for connecting the dots between codes or categories, unveiling relationships and links within our data. Afterward, we used selective coding to distill the complexity of categories we created when using axial coding into a clear, understandable, and coherent story or framework [107]. We coded the results using `Taguette`[3] by using the two authors independently. This analysis helped us understand the definitions of anonymity from the software developer's perspective.

[2] `https://transcribe.com/`
[3] `https://app.taguette.org/`

TABLE 2.1: Demographics results of surveying 34 early career software developers.

Interview Study Question	%
Q1: What is your current professional role?	
Graduate Student	56%
Undergraduate Student	24%
Software Engineer	6%
Site Reliability Engineer	3%
Cloud Infrastructure Engineer	3%
Engineering Manager	3%
Junior Developer	3%
Professor	3%
Q2: What country do you currently reside?	
Canada	91%
France	9%
Q3: How many years of experience do you have?	
0-1 years	49%
1-2 years	15%
2-3 years	21%
3-4 years	9%
4-5 years	3%
More than 5 years	3%
Q4. What is your self-identified gender?	
Male	56%
Female	44%

13 **Data analysis**: The coded data was analyzed to identify recurring themes, patterns, and insights into developers' perceptions of anonymity on social coding platforms.

14 **Reporting results**: Finally, we organized the survey results along with the results from our data analysis and data coding step to report the results in this paper.

2.4 Interview Study Results

We aimed to study early-career developers; therefore, we recruited 34 early-career developers. We use the umbrella term early-career developers to refer to individuals who are graduate and undergraduate students with less than five years of experience in the industry. Early-career developers were chosen because they represent a group that is still in the process of forming professional habits, navigating the challenges of anonymity and privacy, and integrating into the collaborative norms of social coding platforms. Throughout this paper, any reference to "software developers" or "developers" specifically applies to these early-career professionals.

2.4.1 Interview study participants

To characterize participants of the interview study we asked the developers a series of four questions Q1 to Q4 (See Table 2.1).

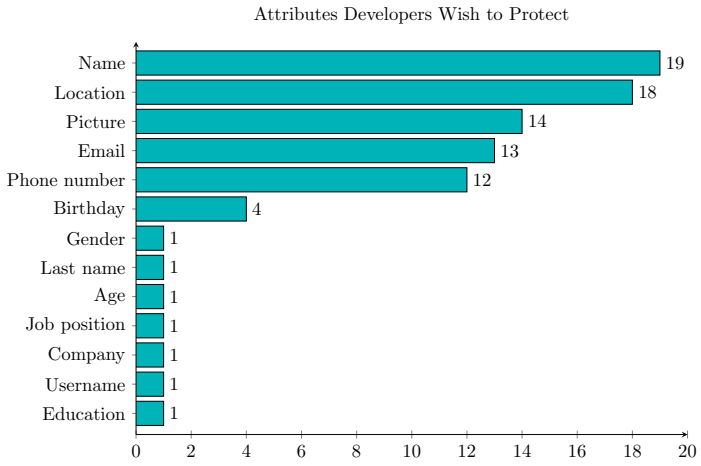


FIGURE 2.2: Attributes important for developers to protect through anonymity (Q6).

Experience and background: The participants in our interview study came from a variety of backgrounds within the software development domain of academia and industry. The largest group, representing 56% (19 out of 34), were Graduate students. Undergraduate students made up another 24% (eight out of 34). The remaining participants were professionals with various titles within the work industry, one Engineering Manager (3%), one Site Reliability Engineer (3%), one Junior Application Developer (3%), and one Cloud Infrastructure Engineer (3%), including two Software Engineers (6%) where one of them was also an Educator (See Table 2.1: Q1). There were 34 people that we interviewed of which 91% of the developers were from North America while 9% were from France, and Europe (See Table 2.1: Q2). Most of the interviewees which accounted for 17 participants had 0-1 years of experience (49%). Followed by five (15%) interviewees who had 1-2 years of experience in the industry. There were seven developers (21%) within the 2-3 year experience range of experience and three developers (9%) with 3-4 years of experience. There was an equal amount of developers (3%) with 4-5 years and more than five years of experience (See Table 2.1: Q3). Out of the total developers we interviewed 56% (19 out of 34) identified as male while 44% (15 out of 34) identified as female (See Table 2.1: Q4).

2.4.2 Definition of anonymity by developers (RQ1)

To answer **RQ1**, we asked the 34 developers a series of six questions from Q5 to Q10 (See Table 2.2) to determine their definition of anonymity.

Definition of anonymity by early career software developers

We asked the developers Q5 (See Table 2.2) to define the term "*being anonymous online*" in their own way, to which they provided a range of answers. While each person individually provided reasons, we identified similarities among these definitions. The majority of developers which is 65% (22 out of 34) defined anonymity as the absence of identifying information such as name, picture, email, phone number, or even linking their accounts across platforms. The exact definition P31 provided was

> *"I think for being anonymous one would probably want to get rid of their profile picture and any personal information, like your name and address. And maybe where you live, and where you work."*

P03 mentioned that to them anonymity meant

> *"If it's just like an empty profile picture, just some random picture, cause then nobody can really tell who you are, not even like your gender. And then the username as well, If it's like some pretty bland like flower underscore one, two, three that could be anybody."*

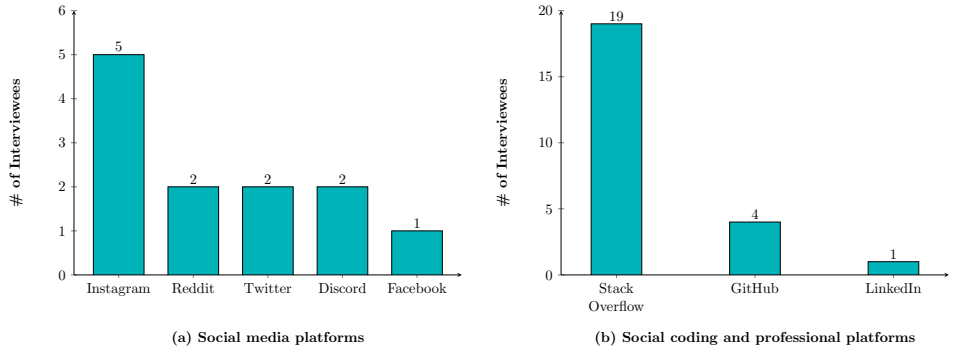
Specifically 9% of the developers (three out of 34) defined anonymity as not having the ability to cross-link identities across platforms. To this end P09 mentioned

> *"I would say that if you can basically not find me from one platform on the other. So on LinkedIn, you can like find me on other platforms, if I want to be known. If I do not want to be known, like, on Stack Overflow, for example, then you are not able to find me there through other platforms using my username, email address, and my picture"*

6% of the developers (two out of 34) defined anonymity as a desire for complete separation between their online and real-life identities. P28 mentioned that

> *"I would say if you have an identity online, and you have identity in real life. You just have no connection between the two identities. So if you don't post the real photo of you on the platform you can't make such a connection."*

This emphasizes a strong preference for hiding information. 6% of the developers (two out of 34) defined anonymity as being completely untraceable through name and



(a) Social media platforms (b) Social coding and professional platforms

FIGURE 2.3: Platforms developers have anonymous accounts (according to their own definition): (a) social media, (b) social coding and professional platforms (Q11).

location searches. An even smaller group, 3% of developers (one out of 34) defined anonymity as having accounts untraceable to their names and pictures, having unverified accounts, or having no activity at all (See Table 2.2).

Key Finding: For early career software developers, anonymity means not sharing identifiable information such as name, picture, location, or professional background, while maintaining a clear separation between one's online and real-life identities and avoiding cross-platform identification.

Profile information developers' protect through anonymity and their reasons

We asked Q6 (See Table 2.2) "*if there were specific profile information (attribute) developers aim to protect with the use of anonymity*". Name (Screenname) topped the list, with 56% (19 out of 34) developers seeking to protect that through anonymity. Similarly, location was a concern for 53% (18 out of 34) developers who preferred anonymity in this aspect. For profile pictures, 41% (14 out of 34) developers wanted anonymity. Developers also mentioned their desire to protect their email addresses, with 38% (13 out of 34) emphasizing the importance of anonymity for this attribute. Phone number anonymity was important to 35% (12 out of 34) developers. 12% (four out of 34) of the developers, expressed concerns about protecting their birthdates. Even less concerning attributes were gender, last name, age, job position, affiliation, username, and education, with only 3% (one out of 34) of developers wanting to protect these through anonymity. (See Figure 2.2).

We followed up with Q7 (see Table 2.2) and asked the developers "*to explain why protecting the specific attributes they mentioned was important to them*". They provided various

reasons for valuing anonymity for these attributes. 21% of the developers (seven out of 34) do not want their online persona to potentially lead to their real-world identities. This notion of separating online and real-world identities is further emphasized by 18% of the developers who wanted to maintain anonymity as a personal preference because they believe that their identifiers like their name and picture are unique to them. Another strong theme is the desire for privacy and avoiding unwanted contact. This is evident in the high percentage (18%) of the developers (six out of 34) who want anonymity to prevent conflicts and unwanted contacts, as well as 9% who want to maintain control over who can identify them in the real world. This might also be linked to a desire not to be confined to a specific online "*bubble of where they work, live and study*" which was mentioned by 9% of the developers. 6% of the developers (two out 34) prioritize not being easily accessible. Other reasons include using their attributes for official accounts and avoiding cross-platform identification to maintain anonymity on their accounts.

Key Finding: The main reason early-career developers want to maintain anonymity is that they do not want their online persona to potentially lead to their real-world identities.

Activities where developers' require anonymity on social coding platforms

We asked the developers Q8 (See Table 2.2) "*if they required anonymity specifically for some activities on the social coding platforms*". The majority of developers 44% (15 out of 34) prioritize anonymity when asking questions on Stack Overflow. A significant concern for 15% of the developers (five out of 34) accounts for the potential fear of negative judgment they might receive based on their online comments on Stack Overflow. This suggests a broader preference for privacy across various online interactions committing to code repositories on GitHub (15%). There were 9% of the developers (three out of 34) who mentioned they require anonymity when browsing, upvoting content (9%), answering questions (9%) or downvoting content (9%).

There were 9% of the developers who mentioned that they do not see the purposes of maintaining anonymity even for certain activities on social coding platforms. Some of the other activities the developers mentioned include working on proprietary software,

TABLE 2.2: Questions and interview survey results of 34 developers with the objective RQ1:"*Definition of anonymity by developers*"

Interview Study Question	%
Q5: How would you describe being anonymous online?	
Not sharing identifying info like name, bio, email, phone number, picture, link to other accounts	65%
No cross-platform identification	9%
Separation of real life and online identities	6%
Unfindable when searched by name and location	6%
No one with any proxy would be able to immediately identify the person based on account information	3%
Not being able to trace back an account to the real person	3%
No activity including commits and posting	3%
Unverified account	3%
Q6: Which profile information are the most important for you to protect your anonymity?	
Name	56%
Location	53%
Picture	41%
Email	38%
Phone number	35%
Birthday	12%
Other information (E.g. gender, age, username, etc)	21%
Q7: Why are these profile information the most important for you to protect your anonymity?	
To avoid narrowing down my profile with more accurate identification	21%
Avoiding conflicts and unwanted contacts	18%
Personal preference because of the uniqueness of one's name	18%
Avoid leading to identify me in the real world	9%
Not being identifiable through google search	9%
Not to be easily accessible	6%
Other reasons (E.g. used for official accounts, avoid cross platform identification through my information, etc)	9%
Q8: For which activities do you find anonymity to be crucial on social coding platforms?	
Asking questions on Stack Overflow	44%
Commenting on Stack Overflow where others can judge me negatively	15%
Committing to repositories on GitHub	15%
Browsing on Stack Overflow	9%
Answering questions on Stack Overflow	9%
Downvoting on Stack Overflow	9%
Upvoting on Stack Overflow	9%
Other activities (E.g. working on proprietary software, posting issues, working on low quality projects , etc)	33%
Q9: Which factors influence your decision to share information or engage with other developers on social coding platforms?	
Quality and substance of developers content	65%
Peer anonymity has influence	21%
Educational background matters	18%
Profile engagement richness matters	18%
Profile richness matters	9%
Other reasons (E.g. post engagement matters, repository engagement matters, etc)	6%
Q10: Do you think that having multiple accounts assists with your anonymity?	
Yes	82%
No	18%

posting issues, working on low quality projects, unintentional anonymity due to limited activity, avoiding heated topics, and not engaging with content at all.

Key Finding: Developers primarily seek anonymity when asking questions (44%), making comments (15%), and committing to repositories (15%) primarily to avoid negative judgment by others on the platform.

Factors influencing developers' decision to share information on social coding platforms

We also asked the developers Q9 (See Table 2.2) *"which factors influence their decisions to share information or engage with other developers on the social coding platforms*". A clear majority of 65% developers (22 out of 34) prioritize the quality and substance of their content. This suggests a preference for valuing contributions based on the quality of

ideas and solutions shared, rather than focusing heavily on individual backgrounds. For 21% of the developers (seven out of 34), the anonymity of their peers on the platform influences their decision to engage with them on the platform.

The richness of profile engagement (interactions with other developers' profiles) is important to 18% of developers (six out of 34) when sharing information or engaging with others. Educational background of the other developers plays a role for about 18% of the developers (six out of 34) to engage with them. Similarly, some of the other factors such as post and repository engagement have a minor influence on the developers decision.

Key Finding: Majority (65%) of developers place a higher value on the quality of content and ideas shared on the platforms rather than on the detailed profiles of contributors when sharing or engaging with other developers on the social coding platforms.

We also asked the developers Q10 (See Table 2.2) "*if they think having multiple accounts can help with anonymity*" where 82% (28 out of 34) developers said yes while 18% (six out of 34) developers said no.

2.4.3 Self-perceived anonymity status (RQ2)

To answer **RQ2**, we asked the developers a series of eight questions Q11 to Q18 (See Table 2.3) to understand their self-perception of anonymity on social media (💬), social coding (</>), and professional platforms (in). To avoid redundancy, we have written the questions we asked during the interview only once and have indicated them with an asterisk (*) to present the results.

Developers' anonymous accounts

* We asked the developers Q11 (See Table 2.3) "*based on their own definition of anonymity, on which social media (💬), social coding (</>), and professional platforms (in) they have anonymous accounts*".

💬 **Social media platforms**: On Instagram 15% (five out of 34) developers reported having anonymous accounts. On Reddit, Twitter, and Discord it was the same number

TABLE 2.3: Questions and interview survey results of 34 developers with the objective of "*RQ2: Self-perceived anonymity status*"

Interview Study Question	%
Q11: Based on your own definition, on which social media, social coding or professional platforms are you anonymous?	
Stack Overflow	56%
Instagram	15%
GitHub	12%
Reddit	6%
Twitter	6%
Discord	6%
Other platforms (E.g. Facebook, LinkedIn)	6%
Q12: Why did you choose to remain anonymous on any of your social media, social coding platforms or professional accounts?	
Do not wish to associate online activity with real-life identity because of the permanency of content online	35%
Peer anonymity influence	9%
Due to the nature of the platform	6%
Not a content creator	6%
Other reasons (E.g. avoid linking questions to me to avoid possible backlash, etc)	54%
Q13: How important is it for you to remain anonymous on social coding platforms?	
See Figure 2.4	
Q14: On which of your social media, social coding or professional platforms do you use real photos?	
LinkedIn	88%
Instagram	68%
Facebook	29%
GitHub	24%
Snapchat	18%
Twitter	18%
Stack Overflow	15%
Youtube	6%
Tiktok	6%
Other platforms (Messenger, Telegram, Discord, etc)	15%
Q15: What social media, social coding, and professional platforms have you included in your resume?	
LinkedIn	94%
GitHub	85%
Q16: In your resume, why have you included, the social media, social coding, professional platforms that you mentioned?	
GitHub	**LinkedIn**
To show case projects 59%	To show past experience 41%
For job recruitment 9%	For networking 24%
To show technical background projects and skills 6%	For job recruitment 9%
To portray professional self 6%	To portray professional self 6%
Other reasons 6%	Other reasons 3%
Q17: On which of your social media, social coding or professional platforms do you have multiple accounts?	
Instagram	36%
GitHub	9%
Twitter	9%
Other platforms (E.g. Youtube, Stack Overflow)	9%
Q18: Why do you have multiple accounts on your social media, social coding and professional platforms you mentioned?	
To showcase hobbies	24%
To separate concerns between close circle and generic follows	15%
To separate concerns between personal circle and work account	6%
To separate concerns between close circle and career development	6%

of people who reported anonymous accounts which was 6% (two out of 34) developers. On Facebook 3% of the developers (one out of 34) reported anonymous accounts.

</> **Social coding platforms**: 32% of the developers (11 out of 34) reported having no anonymous accounts. Majority of the developers 56% (19 out of 34) reported anonymous accounts on at least one platform. From the social coding platforms, Stack Overflow was the platform with the most anonymous accounts reported, with 56% (19 out of 34) anonymous accounts. On GitHub it was 12% (four out of 34) developers who had reported anonymous accounts (See Figure 2.3).

Professional platform: On LinkedIn, it was 3% of the developers who had reported anonymous accounts.

Developers' reasons for having anonymous accounts

* We asked the developers Q12 (See Table 2.3) "*the reason for making an account anonymous on social media (💬), social coding (</>), and professional platforms (in)*".

💬 **Social media platforms**: 9% of the developers (three out of 34) mentioned peer anonymity influence as their reason for maintaining anonymous accounts on Discord. 6% of the developers' (two out of 34) reason for having anonymous accounts on Reddit was because of the nature of the platform which enforces anonymity. 6% of the developers' (two out of 34) reason for maintaining anonymous accounts on Facebook was because they were not content creators. Some of the other reasons the developers mentioned are wish to not be noticed on the platform, having limited activity, to avoid cross platform identification by the platforms, and showcasing hobbies on their social media that they prefer not to display to a broad audience were mentioned by 3% of the developers.

</> **Social coding platforms**: Majority of the developers' 35% (12 out of 34) reason for having anonymous accounts on Stack Overflow is because they do not want to associate their activity on the platform with their real-life identity because they feel once they put out content (in the form of questions, answers, comments) it will remain accessible for a while. To this end, P09 mentioned

> *"I'm anonymous on Stack Overflow because I don't use my real name, there's no picture. I don't want my accounts linked back to me in real life and I don't want that to affect me negatively because of the permanency of the internet so that's why I am anonymous."*

3% of the developers (one out of 34) also have anonymous accounts on Stack Overflow, to avoid negative consequences associated with their online actions or opinions. 3% of the developers were unintentionally anonymous on Stack Overflow due to creating accounts through GitHub. In addition 3% of the developers actively avoid having their online presence tracked across social coding platforms.

in **Professional platform**: Some developers (3%) use LinkedIn anonymously to view profiles.

We conducted a correlation analysis between different platform categories. The average correlation coefficient between social coding platforms (Stack Overflow and GitHub)

TABLE 2.4: Questions and results for 34 developers interview RQ3 objective "*Developer engagement with privacy policies*"

Interview Study Question	
Q19: Which of the privacy policies, either Stack Overflow's or GitHub's, have you read?	**%**
Stack Overflow	0%
GitHub	0%
Q20: Why you have not read these privacy policies?	**%**
Lengthy	71%
Trust in platforms	9%
Need for use of the app	9%
Other reasons	18%
Q21: How much time would it take you to read Stack Overflow privacy policy?	**minutes**
Average time	27 min.
Q22: How likely is it for you to indeed read through Stack Overflow's policy fully?	**Likert**
See Figure 2.8	
Q23: What was the most important policy in Stack Overflow's privacy policy?	**%**
"collection of information such as your username"	37%
Q24: How much time would it take you to read GitHub privacy policy?	**minutes**
Average time	12 min.
Q25: How likely is it for you to indeed read through GitHub's policy fully?	**Likert**
See Figure 2.8	
Q26: What was the most important policy in GitHub's privacy policy?	**%**
"collection of information such as your name..."	39%

and social media platforms (Instagram, Facebook, Reddit, Twitter, Discord) is 0.65, indicating a strong correlation. This suggests that developers who have anonymous accounts on social coding platforms are also likely to behave the same on social media. On the flip side, the average correlation coefficient between the developers on social coding platforms and professional platforms (LinkedIn) is 0.34, which indicates a weaker correlation. This implies that the developers presence on social coding platforms does not strongly correlate with their presence on professional platforms. The average correlation coefficient between social media platforms and professional platforms is 0.43. This suggests that developers with anonymous accounts on social media have more of a tendency of also behaving the same on professional platforms.

Key Finding: When comparing the developer's behavior in terms of anonymous accounts on social media platforms against social coding platforms and professional platforms we see that developers tend to have anonymous accounts on social coding platforms with a high correlation of 0.65 to avoid associating their activity on the platform with their real-life identity.

Importance of anonymity on social coding platforms

We asked the developers Q13 (See Table 2.3) *"to express the importance they place for anonymity on Stack Overflow and GitHub"*. Their response was reported on a Likert scale from one (very low importance) to five (very high importance).

On Stack Overflow seven developers (21%) reported a very low importance for anonymity, seven developers (21%) indicated a low importance for anonymity, 11 developers (32%) felt a moderate importance for anonymity, six developers (18%) reported a high importance for anonymity, three developers (9%) indicated a very high importance for anonymity.

On GitHub, 10 developers (29%) reported a very low importance for anonymity, 11 developers (32%) indicated a low importance for anonymity, eight developers (24%) felt a moderate importance for anonymity, four developers (12%) reported a high importance for anonymity and one developer (3%) indicated a very high importance for anonymity (See Table 2.3: Q13).

Key Finding: Developers placed a higher perceived importance on anonymity on Stack Overflow compared to GitHub in their early careers, primarily due to fears of having their questions attributed to them and the potential negative impact on future employment opportunities.

Developers' choice of real profile pictures

* We asked the developers Q14 (See Table 2.3) *"on which on social media (💬), social coding (</>), and professional platforms (in) they use real profile pictures"*. We define a real profile picture as one that contains a graphic representation of the account owner.

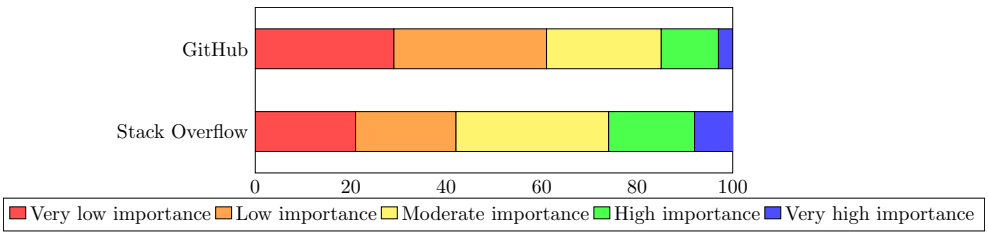


FIGURE 2.4: Importance of anonymity for the developers on Stack Overflow and GitHub (Q13).

💬 **Social media platforms**: On Instagram, 68% of the developers (23 out of 34) use real photos. On Facebook, 29% of the developers (10 out of 34) use their real photos. On Snapchat and Twitter, 18% (six out of 34) use real photos. On YouTube and TikTok, 6% (two out of 34) use real profile pictures. 6% of the developers reported not using real photos on any platform. On other social media like Messenger, iMessage, Discord, Bereal, and Telegram had 3% each using real photos (See Figure 2.5).

</> **Social coding platforms**: On GitHub 24% (eight out of 34) developers use their real photos while on Stack Overflow 15% (five out of 34) developers use their real photo.

Professional platforms: Our interview study reveals that the developers have a clear preference for using real profile pictures on professional platforms. The developers more frequently share their profile picture on LinkedIn, with 88% (30 out of 34) using their real profile pictures.

We conducted a correlation analysis regarding the tendency to share real profile pictures across different platform categories. The average correlation coefficient between social coding platforms (Stack Overflow and GitHub) and social media platforms is -0.18, which indicates a weak negative correlation. This suggests that individuals who share real profile pictures on social coding platforms are less likely to do so on social media. In contrast, the average correlation coefficient between social coding platforms and professional platforms (LinkedIn) is 0.21, reflecting a weak positive correlation, suggesting a slight tendency for individuals who share real profile pictures on social coding platforms to also do so on professional platforms. The strongest correlation we observed is between social media platforms and professional platforms, with an average correlation coefficient of 0.68. This indicates a strong positive correlation, suggesting that individuals who share real profile pictures on social media are highly likely to do the same on professional platforms like LinkedIn.

Key Finding: When comparing developers' tendencies to share profile pictures on social media versus social coding and professional platforms, we can see that they are less likely to share profile pictures on social coding platforms with a correlation of -0.18 and more likely to use personal identification on professional and social media platforms with a correlation of 0.68.

Platforms developers include in their resume with reasons

* We asked the developers Q15 (See Table 2.3) *"which social media (💬), social coding (</>), and professional platforms (in) they provide a link to in their resume"*. We found that none of the developers provided their social media in their resumes.

</> **Social coding platforms**: 85% of the developers (29 out of 34) include GitHub in their resumes. We followed up with Q16 (See Table 2.3) and asked the developers *"why they included the social coding platform in their resume"*. The reason 59% of the developers included GitHub in their resume was to showcase projects. To this end, P23 stated:

> *"I have included GitHub in my resume because it's more professional and I use them more to showcase my projects."*

While 9% of the developers mentioned job recruitment as a reason. 6% of the developers mentioned demonstrating technical skills as a reason. 6% of the developers incorporate GitHub in their resume to portray their professional selves.

in **Professional platforms**: We found that 94% of the developers (32 out of 34) include LinkedIn in their resume.

We followed up with Q16 (See Table 2.3) and asked the developers *"why they included the professional platform in their resume"*. 41% of the developers mentioned including LinkedIn in their resume to highlight past experience. The reason mentioned by 24% of the developers (eight out of 34) was to focus more on professional networking. While 9% of the developers' (three out of 34) reason was for job recruitment. Only 3% of the

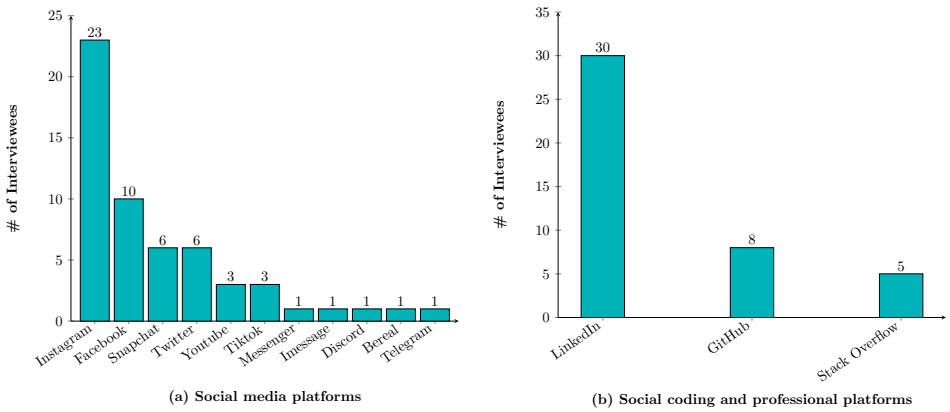


FIGURE 2.5: Platforms developers use real profile pictures on: (a) social media, (b) social coding and professional platforms (Q14).

TABLE 2.5: Important statements for the developers from Stack Overflow's privacy policy (Q23).

Stack Overflow's privacy policy	% of people
We collect information such as your username, password, email address, and IP address (collected at each site visit). This information is also used to identify individual users.	37%
Your username and additional content you provide, including your profile picture, question and answer content, and your reputation, are visible publicly.	19%
We do not and will not sell your personal information	15%
We collect information through third-party services, such as Third party services Google Ads and Google Analytics, for the purposes of serving targeted advertisements, personalized content, and analytics.	15%
We collect location information about **non-members** including your IP address, your location, browser information, and how you came to the Stack Overflow Network for individuals who are non-members to provide advertising, localized content, recommendations, and marketing.	9%
We collect location information about **members** including your IP address, your location, browser information, and how you came to the Stack Overflow Network for individuals who are non-members to provide advertising, localized content, recommendations, and marketing.	6%
We collect contact data including delivery address, email address, and telephone numbers for operational purposes, including providing Services and Products.	3%
We use web browser and mobile analytics to allow us to understand Network and Apps functionality. In doing so, we record information including, for example, how often you visit the Network, how often you contribute content, Network and App performance data	3%

developers (one out of 34) use LinkedIn for self-promotion. Our findings suggest developers rely more on established platforms to include in their resume, with GitHub ideal for code and projects (85%), and LinkedIn for professional presentation and networking (94%). The reason developers share their GitHub and LinkedIn profiles in their resume is to showcase projects (59%) and to show past experience (41%).

Developers' multiple accounts along with their reasons

* In Q17 (See Table 2.3), we asked the developers "*on which on social media (💬), social coding (</>), and professional platforms (in) they have multiple accounts*".

💬 **Social media platforms**: The interview survey revealed that the majority of developers 50% (17 out of 34) had a single account across platforms. On Instagram 36% (12 out of 34) developers have multiple accounts. On Twitter 9% (three out of 34) developers (See Figure 2.6). We followed up with Q18 (See Table 2.3) "*the reason as to why the developers have multiple accounts on social media platforms*". Some developers mentioned having a separate account specifically for showcasing hobbies on Instagram which was 24% (eight out of 34). In addition on Instagram where 15% developers (five out of 34) reported using multiple accounts to separate close friends from general followers. To this end P11 stated:

> *"I have multiple accounts on Instagram for my different hobbies, which I just don't really want to continue."*

TABLE 2.6: Important statements for the developers from GitHub's privacy policy (Q25).

GitHub's privacy policy	% of people
We collect information provided for promotional communications such as name, email address, and company name.	39%
We may share Personal Data with the following recipients: Abuse and Fraud Prevention Entities. We may disclose Personal Data based on a good faith belief it is needed to prevent fraud, abuse, or attacks on our Services, or to protect the safety of GitHub and our users.	30%
When you use our Services, we collect Personal Data included as part of the information you provide such as code, inputs, text, documents, images, or feedback.	24%
We use cookies and similar technologies to provide essential functionality like storing settings and recognizing you while using our Services.	9%
We collect information to create a user profile, which may include a photo, additional email addresses, job title, or biography.	6%
GitHub uses cookies for users login information and preferences to provide, secure and improve our Service or to develop new features and functionality of our Service.	6%
In some cases, you provide us with ethnicity, gender, or similar demographic details which we collect.	3%

On Twitter 6% of the developers (two out of 34) mentioned using multiple accounts to separate close circle and career development. This allows them to maintain privacy boundaries, address specific needs on different platforms, and potentially maintain some degree of anonymity in certain contexts.

💬 **Social coding platforms**: 9% of the developers (three out of 34) on GitHub had multiple accounts. While on Stack Overflow 3% of developers (one out of 34) had multiple accounts. We followed up with Q18 (See Table 2.3) "*the reason as to why the developers have multiple accounts on social coding platforms*". Developers who had multiple accounts mentioned they wished to maintain clear boundaries between their professional and personal online identities on GitHub. 6% of the developers (two out of 34) mentioned that they have multiple accounts on GitHub for potentially managing personal projects or work-related content.

Key Finding: When comparing developers' attitudes on social coding and professional platforms to their social media, we observe a tendency to maintain multiple accounts on social media. This is driven by their desire to distinguish between different types of connections.

2.4.4 Developers' engagement with privacy policies (RQ3)

To answer **RQ3** we asked the developers a series of eight questions to determine their engagement with privacy policies.

Previous reading status of social coding platforms privacy policies

We asked the developers Q19 (See Table 2.4) "*if they had read the privacy policies of Stack Overflow or GitHub*" to which 100% (34 out of 34) developers said that they had not done so. We also asked Q20 (See Table 2.4), "*the reasons as to why they did not read the privacy policies of Stack Overflow and GitHub*" to which they listed several reasons (See Table 2.4).

The most significant factor is the lengthiness of the policies which is mentioned as a reason to skip privacy policies due to the high volume of information by 71% of the developers. Some developers (3%) stated that the lack of alternatives even if they disagree with the policies is the reason they do not read privacy policies. 3% of the developers consider the policies similar to others they've encountered that discourages reading. 12% of the developers trust the platforms to handle their data responsibly, while some 9% of them see using the app as a necessity despite potential privacy concerns.

Key Finding: We found that on average 100% of the early-career developers did not read privacy policies of Stack Overflow and GitHub. The main reason is the lengthiness of documents (71%), trust in the platforms (12%), and the need to use them (9%).

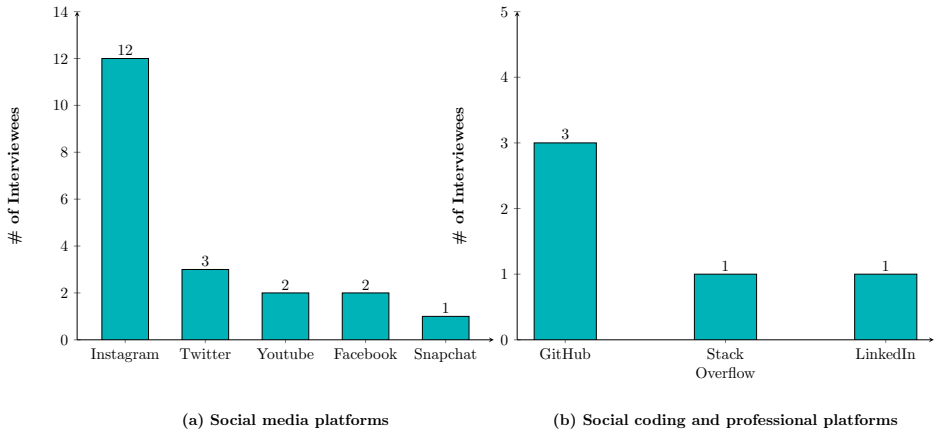


FIGURE 2.6: Developers have multiple accounts on: (a) Social media platforms, (b) Social coding and professional platforms (Q17).

Average time perceived to be taken by Developers to read social coding platforms' privacy policies

We asked the developers Q21 (See Table 2.4), "*how much time they believed it would take them to read the privacy policy of Stack Overflow*". Stack Overflow's privacy policy, which was coded and addressed under personal data was 12 pages long. The developers said on average it would take them 27 minutes to read due to the length. We asked the developers Q24 (See Table 2.4), "*how much time they believed it would take them to read the privacy policy of GitHub*". GitHub's privacy policy, which was coded and addressed under personal data was five pages, to which on average the developers said it would take them 12 minutes to read due to the length.

Key Finding: We found developers rarely read privacy policies on Stack Overflow or GitHub in their own time because it would take them on average 27 and 12 minutes respectively to read in full.

Current reading status of privacy policies

In our controlled study, we set specific reading times for developers to review the privacy policies during the interview. Developers were given 10 minutes to read Stack Overflow's 12-page privacy policy and 5 minutes for GitHub's five-page privacy policy. These times were determined based on an estimated average reading speed of one minute per page [40].

After reading the privacy policies we asked the developers Q22 (See Table 2.4) which is "*how likely it would be for them to read the privacy policy of Stack Overflow after knowing*

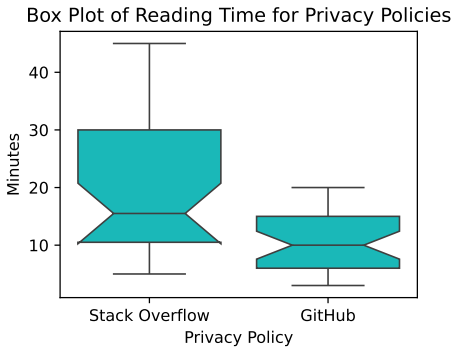


FIGURE 2.7: Likeliness of reading Stack Overflow' and GitHub's privacy policy (Q22) and (Q25).

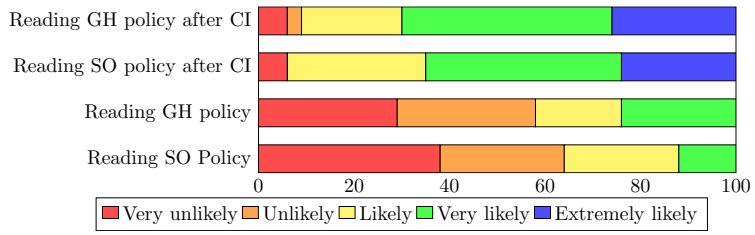


FIGURE 2.8: Likeliness of reading Stack Overflow's and GitHub's privacy policy before and after applying Contextual Integrity on the policy (Q22), (Q25), (Q30), and (Q31).

the terminologies". This question is a five-point Likert scale which ranges from very unlikely (score of one) to extremely likely (score of five). On Stack Overflow the majority of developers, 38% (13 out of 34) developers reported a very unlikely chance of reading the privacy policy (score of one). This was followed by 26% (nine out of 34) indicating an unlikely chance of reading the policy (See Figure 2.8). This data suggests a need for increased awareness among developers about the importance of understanding Stack Overflow's privacy practices.

When we asked the developers Q25 (See Table 2.4) "*how likely it would be for them to read the privacy policy of GitHub after knowing the terminologies*". We observed that 29% (10 out of 34) developers reported a very unlikely chance of reading it (score of one), the same percentage (29% or 10 out of 34) indicated a somewhat unlikely chance (score of two) (See Figure 2.8). This data suggests a need for improved communication and education about GitHub's privacy practices.

Key Finding: After the developers read the privacy policies, the majority reported that they are either very unlikely or unlikely on the Likert scale to read the privacy policies of Stack Overflow and GitHub (64% and 58%, respectively). Our interview results suggest a need for improved communication and education around Stack Overflow's and GitHub's privacy practices.

Important statements from the privacy policies of social coding platforms

After reading the privacy policy of Stack Overflow we asked each developer Q23 (See Table 2.4) "*to point to an important statement they found*". They mentioned nine statements overall which they said were important to them (See Table 2.5). One of the most important statements for 37% of the developers (12 out of 34) was

TABLE 2.7: Questions and results for 34 developers interview RQ4 objective "*Impact of context on privacy policy understanding*"

Interview Study Question	
Q27: How would you rank Stack Overflow's statements based on CI in order of importance?	**See Figure: 2.9**
Q28: How would you rank GitHub's statements based on CI in order of importance?	**See Figure: 2.10**
Q29: Do you believe adding the CI structure helped you with understanding of these policies?	**%**
Yes	97%
No	3%
Q30: How likely is it for you to read Stack Overflow's policy if it was restructured as such?	**See Figure: 2.8**
Q31: How likely is it for you to read GitHub's policy if it was restructured as such?	**See Figure: 2.8**

> *"We collect information such as your username, password, email address, IP address (collected at each site visit) in which this information is also used to identify individual users."*

The statement that was second most important to 19% of the developers (six out of 24) on Stack Overflow is

> *"Your username and additional content you provide including your profile picture, question and answer content and your reputation are visible publicly."*

The third most important statement to 15% of the developers (five out of 24) on Stack Overflow is

> *"We do not and will not sell your personal information."*

After reading the privacy policy of GitHub we similarly asked each developer Q26 (See Table 2.4) "*to point to an important statement they found*". They mentioned eight statements that they said were important to them (See Table 2.6). One of the most important statements for the developers which were identified by 39% (13 out of 34) of the developers was

> *"We collect information provided for promotional communications such as name, email address, and company name."*

The second most important statement to 30% of the developers (10 out of 34) on GitHub is

> *"When you use our Services, we collect Personal Data included as part of the information you provide such as code, inputs, text, documents, images, or feedback.*

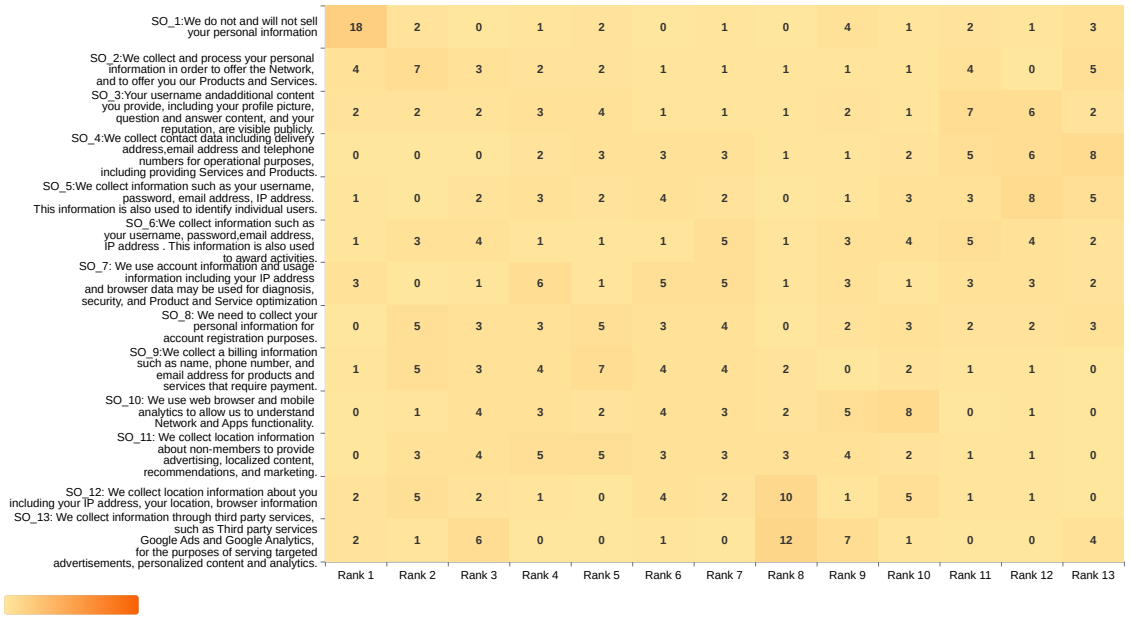


FIGURE 2.9: Heat map of the Stack Overflow privacy policy ranking provided by the developers for the documents that were restructured using Contextual Integrity

The statement that was the third most important to 24% of the developers (eight out of 34) on GitHub is

> *"When you use our Services, we collect Personal Data included as part of the information you provide such as code, inputs, text, documents, images, or feedback."*

2.4.5 Impact of context on privacy policy understanding (RQ4)

Contextual Integrity is a theory that views privacy as the appropriate flow of information, determined by compliance to established and legitimate informational norms within a specific social context [81]. We used the framework proposed by contextual integrity for the privacy policies to recognize appropriate information flows on Stack Overflow and GitHub. We used the sender, subject, recipient, attribute, and transmission principle to form these complete information flows [99]. To answer **RQ4** we asked the developers a series of five questions by adjusting the privacy policies of Stack Overflow and GitHub. We adjusted the 12-page Stack Overflow and five-page GitHub privacy policies using the framework Contextual Integrity (CI) [82].

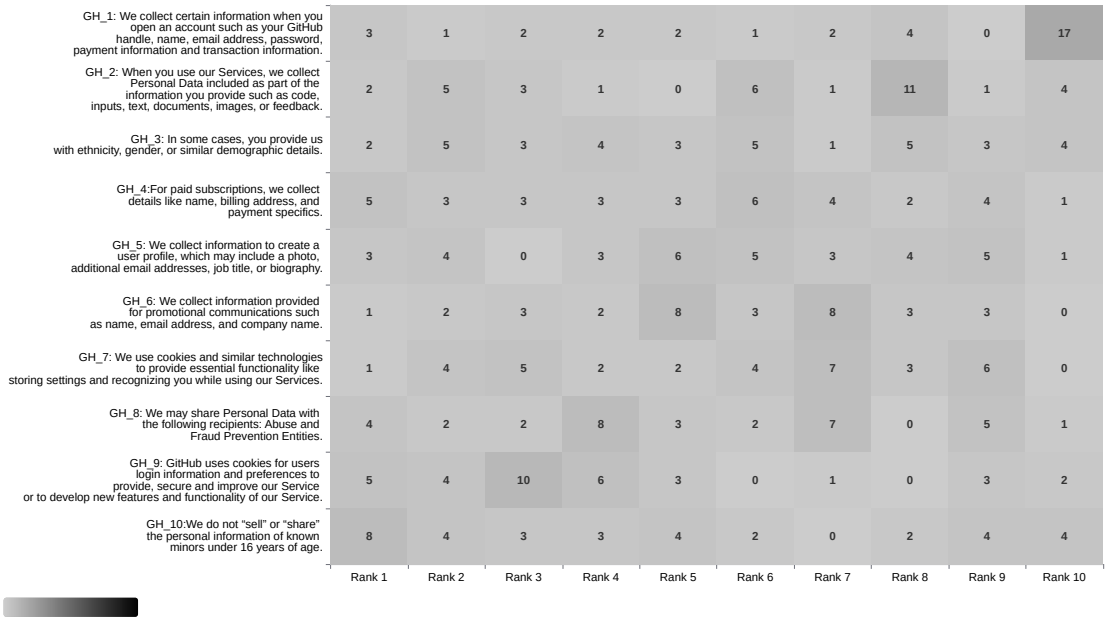


FIGURE 2.10: Heat map of the GitHub privacy policy ranking provided by the developers for the documents that were restructured using Contextual Integrity

Ranking using Contextual integrity on social coding platforms privacy policy

In our research, we aimed to understand how developers perceive privacy policies when they are presented using the theory of contextual integrity. Contextual integrity, as a framework, emphasizes the importance of context in understanding privacy norms and expectations. We aimed to see if restructuring privacy policies according to this framework would impact developers' understanding of these policies.

To achieve this, we extracted complete information flows from the privacy policies of Stack Overflow and GitHub. The research team restructured these policies according to the principles of contextual integrity, ensuring that the context of information flow was clearly articulated. To this end we extracted the statements and clearly listed out the sender, subject, receiver, attribute, transmission flow from the privacy policies. For Stack Overflow, we identified 13 information flow statements from the restructured privacy policy. We presented these statements alongside the original versions to the developers and asked them to rank these statements in terms of importance. This was done through Q27 (See Table 2.7), which we asked the developers to "*rank the statements of Stack Overflow after applying contextual integrity*."

Our results showed a strong preference for certain statements. The statement "*We do not sell and will not sell your personal information*" was particularly noteworthy, as it was ranked first 18 times by the developers (See Figure 2.9). This indicates a high level of concern among developers about the sale of personal information, highlighting their priority for policies that protect their data from commercial exploitation.

Similarly, for GitHub, we identified 10 complete information flow statements from the restructured privacy policy. Developers were asked to rank these statements through Q28 (See Table 2.7), where we asked "*to rank the statements of GitHub after applying contextual integrity*." In the case of GitHub, the statement "*We do not sell and will not sell the personal information of known minors under 16 years of age*" emerged as the most significant, being ranked first eight times by the developers (See Figure See Figure 2.10)). This reflects a strong concern for the protection of minors' personal information, which aligns with the general societal expectations about protecting vulnerable populations.

These findings suggest that when privacy policies are presented with clear contextual integrity, certain statements resonate more strongly with developers. These statements often relate to the non-commercial use of personal information and the protection of vulnerable groups. The implications of these results are significant for the software developers on the platform and policymakers. By structuring privacy policies in a way that aligns with the principles of contextual integrity, platforms can enhance transparency and trust among their users. This can lead to higher engagement and a more positive perception of the platform's commitment to user privacy.

Impact of Contextual integrity on developer understanding

We asked the developers Q29 (See Table 2.7) which is *"if they believed that changing the privacy policy using Contextual Integrity helped to understand Stack Overflow's and GitHub's policy better"* to which all 97% (33 out of 34) developers said yes. In addition, when we asked them Q30 *"about the likeliness of reading the restructured privacy policies on Stack Overflow"*. 6% of the developers (two out of 34) reported a very unlikely chance of reading a restructured privacy policy while 29% of the developers (10 out of 34) expressed a likely chance of reading a restructured Stack Overflow privacy policy. There were 41% of the developers (14 out of 34) that indicated a very likely chance of reading the restructured policy. While 24% of the developers (eight out of 34) reported an extremely likely chance

of reading a restructured Stack Overflow privacy policy (See Figure 2.8). Our findings suggest that the traditional presentation of privacy policies lacks effective alternatives for enhancing comprehension. Restructuring privacy policies using Contextual Integrity principles can significantly improve users' understanding and engagement with these documents.

We also asked the developers Q31 (See Table 2.7) which is *"about the likeliness of reading the restructured privacy policies of GitHub"*. On GitHub 6% of the developers (two out of 34) reported a very unlikely chance of reading a restructured privacy policy, while 3% of the developers (one out of 34) indicated an unlikely chance of reading the restructured policy. There were 21% of the developers (seven out of 34) who expressed a likely chance of reading a restructured privacy policy while 44% of the developers (15 out of 34) indicated a very likely chance of reading the restructured policy. There were 26% of the developers (nine out of 34) reported an extremely likely chance of reading a restructured GitHub privacy policy (See Figure 2.8).

A comparison of developers' likelihood of reading restructured privacy policies on Stack Overflow and GitHub highlights both differences and similarities. For both platforms, 6% of the developers reported a very unlikely chance of reading the restructured privacy policy. However, there is a difference in the unlikely category: 29% of developers indicated an unlikely chance for Stack Overflow, compared to 24% for GitHub. In the likely category, 24% of developers expressed that they would read the restructured policy for Stack Overflow, while 15% did so for GitHub. The developers on GitHub reported a higher likelihood of 26% indicating a very likely chance of reading the restructured policy compared to 24% for Stack Overflow. Additionally, the percentage of developers reporting an extremely likely chance was higher for GitHub (15%) than for Stack Overflow (12%). These results indicate that while the initial disinterest (very unlikely) is the same for both platforms, there are differences in the levels of interest beyond that, with GitHub showing slightly higher engagement in the more positive likelihood categories.

Key Finding: We found using Contextual Integrity resulted in a much higher likelihood of developers saying they would read the privacy policies in the future compared to the format. We observed an average of 22% increase in the developers' extremely likely chance of reading the restructured policies of Stack Overflow and GitHub.

2.5 Discussions and Implications

In this section, we discuss our findings, explore their implications, and outline directions for future research.

2.5.1 Impact of intersectional diversity factors on the urge for anonymity

Intersectional diversity acknowledges that we must consider everything that can marginalize people – gender, race, class, sexual orientation, physical ability, and that these forms of oppression can overlap [22]. There has been research done on perceived intersectional diversity and its impact on the software engineering field [90]. Diversity can emerge from different factors such as gender, nationality, age, and race [90].

We collected the developer's race, sexual orientation, indigenous identity, and disability identity. From the 34 developers, we interviewed 94% (32 out of 34) of the developers identified as Heterosexual while 3% of them identified as asexual and the rest preferred not to answer. Out of the 34 survey participants, the largest racial identity the developers identified with was Asian, making up approximately 62% (21 out of 34) of the developers. Middle Eastern or North African respondents made up 18% (six out of 34), followed by Black or African Americans at 6% (two out of 34), and White at 12% (four out of 34). In terms of indigenous identity, all of the developers did not identify as indigenous. In terms of disability identity 3% of the developers identified as having a disability while 97% of the developers did not. We did not find any significant differences between the group of developers who participated in the interview.

2.5.2 Does Gender Identity have an Impact on Anonymity?

To analyze the relation of gender with anonymity we performed a Mann-Whitney U test on the male and female groups. In addition, we calculated the proportions of the developers who had perceived anonymous accounts as well as perceived multiple accounts. We obtained a p-value of 0.040, which is below the common significance threshold of 0.05. This finding allows us to reject the null hypothesis indicating a statistically significant difference in the anonymity responses between the two groups. Hence, this test suggests that males and females differ significantly in their self-perception of anonymity. The Mann-Whitney U test does not specify the exact nature of the difference such as whether males are more or less anonymous than females. The Mann-Whitney U test does indicate that their responses regarding anonymity are not distributed in the same way. To see the distribution difference we calculate the proportions. From the male interviewees, 52.63% consider themselves anonymous, while 47.37% do not (See Table 2.9). In contrast, 87.01.% of the female interviewees report perceived anonymity, with only 12.99% indicating otherwise. These findings reflect a clear disparity in how anonymity is perceived across genders where the female interviewees have more perceived anonymous accounts when compared to the male interviewees.

We analyzed the relationship between gender and multiple accounts. The Mann-Whitney U test p-value (0.0002) indicates that there is a statistically significant difference between the proportions of males and females regarding the presence of multiple accounts. Since this p-value is well below the standard threshold for significance (See Table 2.9). We can reject the null hypothesis by concluding that the differences in the distribution of multiple accounts between males and females are significant. There is a significant difference in the tendency to have multiple accounts between males and females. This could imply differences in behavior or preferences regarding social media or coding platforms based on gender. The proportions reveal big differences between male and female early-career developers regarding multiple accounts on social media and social coding platforms. For males, the data shows that 78.95% do not have multiple accounts, while only 21.05% report having at least one (See Table 2.9). This indicates that a significant majority of males prefer to maintain a single account. In comparison, the findings for females show that a larger percentage, 86.67%, report having multiple accounts, with only

TABLE 2.8: Stack Overflow activity metric with the definition of the metric.

Stack Overflow Metric	Description
`Questions`	# of questions asked by the user
`Answers`	# of answers provided by the user
`UpVotes`	# of UpVotes cast by the user to posts of other users
`DownVotes`	# of DownVotes cast by the user to posts of other users
`QuestionsViewed`	# of times the user's questions were viewed by others
`Favorites`	# of times the user has received interesting tags for their questions
`AccountCreated`	# of days since the account was created

13.33% indicating they do not. This highlights a big difference in account usage between the genders, suggesting that females are more likely to engage in the practice of having multiple accounts.

2.5.3 Perception of Anonymity Compared to Reality

One of the key insights from our interviews was that developers had differing perceptions of their own anonymity on social coding platforms. Specifically, 55% of the participants believed they were anonymous on Stack Overflow, while 12% felt the same about GitHub. This perceived anonymity often arose from developers using pseudonyms or incomplete profiles, which gave them the impression that they were not easily identifiable. To test these perceptions against reality, we conducted a manual analysis of the developers' profile attributes on both platforms to assess how much identifiable information was publicly available (See Table 2.10). This included reviewing whether they completed their profile dashboards by sharing personal details (eg. name, location), or used consistent usernames.

The results revealed a significant gap between perception and reality. On Stack Overflow, only 18% of the developers were truly anonymous, compared to the 55% who believed they were. Similarly, on GitHub, the actual percentage of anonymous developers was 9%, in comparison to the 12% who perceived themselves as anonymous. This disconnect between perceived and actual anonymity highlights how developers may overestimate their level of privacy on these platforms. While they may believe they are operating anonymously, the interconnected nature of online profiles and the ease of cross-referencing data between different platforms can make them more identifiable than they realize. This finding highlights the importance of understanding how easily information shared across platforms can be combined to breach perceived anonymity.

TABLE 2.9: Percentage of anonymous accounts and multiple accounts of the developers with gender

Group	Proportion
Male	Anonymous Accounts: 52.63% Multiple Accounts: 21.05%
Female	Anonymous Accounts: 87.01% Multiple Accounts: 86.67%
Mann-Whitney U p-value	Anonymous Accounts: 0.04 Multiple Accounts: 0.0002

2.5.4 Triangulating Developer Activity and Profile Completeness on Stack Overflow

We had a list of 34 users from our interview study participants who used both Stack Overflow and GitHub. At the end of the interview, we asked them if they consented with to share their Stack Overflow and GitHub accounts. Out of the total 34 developers 18 of the developers had active Stack Overflow accounts and 32 of the developers had active GitHub accounts. We measured their activity data on both platforms to see if it had a correlation with their Profile availability. We followed a previous paper done by Badashian et al. [7] to measure activity data done by Stack Overflow and GitHub users. To measure the Stack Overflow users' activity data we used `BigQuery`[4] which had the users all time activity data. The activity metrics we measured for Stack Overflow are the number of questions, number of answers, total downvotes received on posts, total upvotes received on posts, total questions viewed, total favorites, number of profile views, and days since the account was created. A full explanation of the attributes along with their definition is presented in Table 2.8.

To explore the potential connection between a developer's profile completeness on Stack Overflow and their activity on the platform, we conducted a correlation analysis. We focused on eleven profile attributes provided by Stack Overflow's user dashboard. These attributes included links to Twitter, GitHub, a personal website, Stack Exchange, Meta Stack Exchange, location information, avatar, username, user ID, and Stack Overflow user link itself.

[4]https://console.cloud.google.com/bigquery

TABLE 2.10: Perceived vs. Actual Anonymity on Social Coding Platforms

Platform	Perceived Anonymity (%)	Actual Anonymity (%)
Stack Overflow	55%	18%
GitHub	12%	9%

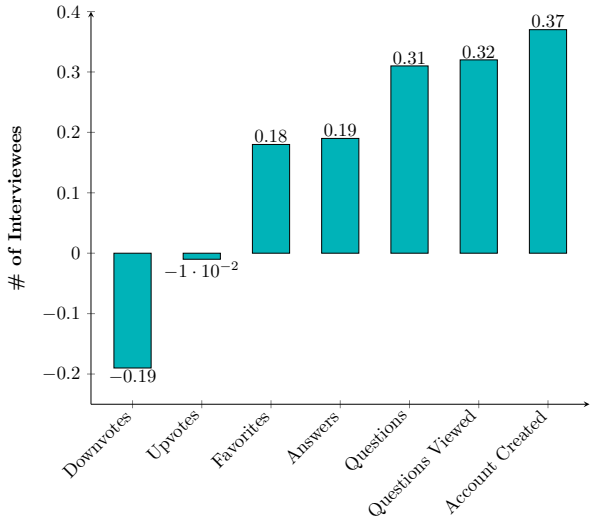


FIGURE 2.11: Stack Overflow activity metric with Stack Overflow attribute availability.

We created a profile completeness score for each user by calculating the number of attributes they had filled out. This score ranged from zero (no attributes) to 11 (all attributes). Next, we examined the correlation between this profile completeness score and various activity metrics on Stack Overflow. These activity metrics included questions asked, answers provided, upvotes and downvotes, questions viewed, favorited posts, and account creation date. Our analysis revealed a positive correlation between the completeness of Stack Overflow profiles and various user activity metrics. Users with more complete profiles (as measured by the number of profile attributes filled out) tend to have higher engagement on the platform. A breakdown of the correlation coefficients between profile completeness and specific activity metrics is presented in Figure 2.11.

2.5.5 Triangulating Developer Activity and Profile Completeness on GitHub

For the developers from GitHub we used the metrics number of commits, number of pull requests, number of repos, number of followers, number of followings, and date on

TABLE 2.11: GitHub activity metric with the definition of the metrics.

GitHub Metric	Description
`Commits`	# of commits made by the developer.
`PullReqs`	# of times the developer notified others about their changes so that they can apply them if they wish to do so
`IssuesReported`	# of issues the developer reported.
`Repository`	# of public repositories of the developer
`Followers`	# of people who follow the developer.
`Following`	# of people the developer is following
`AccountCreated`	The time span since the account was created

when the account was created. We used these metrics because we followed the approach of Badashian et al. [7]. We have described the metrics we used according to the paper by Badashian et al. (See Table 2.11). We retrieved their all time metrics of these developers using the `GitHub Search API`[5].

We conducted a correlation analysis to explore the potential connection between a developer's profile completeness on GitHub and their activity on the platform. We focused on twelve attributes available on a GitHub user profile, including login name, screen name, node ID, avatar, username, company, blog link, location, email (if publicly available), hireability status, bio, and Twitter username.

For each user, we created a profile completeness score based on the number of these attributes they had filled out. This score ranged from zero (no attributes) to 12 (all attributes). Next, we examined the correlation between this profile completeness score and various GitHub activity metrics. These metrics included the number of commits, pull requests submitted, issues reported, repositories created, followers, and following users, along with the account creation date. The result of the correlation coefficients between Correlation between GitHub Profile availability and the metrics (See Figure 2.12).

Our findings show there is a positive correlation between the completeness of developer profiles on Stack Overflow and GitHub and their activity on those platforms. In other words, developers who fill out more information in their profiles tend to be more active on the platform (asking/answering questions on Stack Overflow, or committing code/creating repositories on GitHub).

2.6 Threats to Validity

In this section of the paper, we will describe the threats to the validity of the research as described by Wohlin et al. [119, 118, 32, 4, 54].

[5]https://docs.github.com/en/rest/search/search

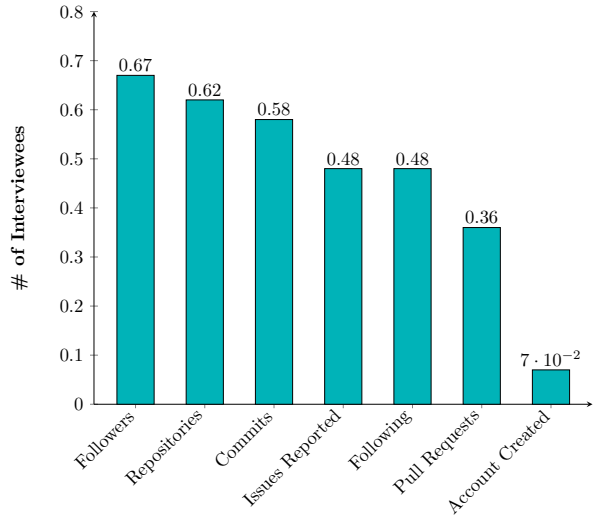


FIGURE 2.12: GitHub activity metric correlation with GitHub attribute availability.

2.6.1 Conclusion Validity

Our study's conclusion validity concerns whether the relationship between treatment and outcome is accurately drawn. By interviewing 34 developers, we aimed to understand their perspectives on anonymity and engagement on social coding platforms. Given the diversity of responses and the consistency in prioritizing content quality over detailed profiles, we believe our conclusions accurately reflect the developers' viewpoints. However, it's important to note that the sample size and demographic factors may still influence the generalizability of our findings.

2.6.2 Construct Validity

Construct validity assesses whether we are measuring the right things. In this study, we designed our interview questions to capture developers' definitions and perceptions of anonymity, as well as their engagement with privacy policies. By directly asking developers about their experiences and motivations, we ensured that our objectives were aligned with the research questions. The interview format allowed for in-depth exploration of these topics, bettering the validity of our constructs.

2.6.3 Internal Validity

Internal validity examines if the treatment caused the outcome. In our study, we investigated whether presenting privacy policies with contextual integrity (CI) principles impacted developers' understanding. The results indicate a preference for content quality and anonymity. However, while we observed this preference, it is challenging to definitely attribute it only to the CI principles without considering other influencing factors. Further controlled experiments could help isolate the specific impact of CI principles.

2.6.4 External Validity

External validity examines if the generalizability of the results go beyond this study. Our findings, based on interviews with 34 developers, provide valuable insights into the preference for anonymity. However, the results may be influenced by the specific demographic and professional backgrounds of our participants since we used snowball sampling to select the participants. Additional studies involving a broader and more diverse developer population are needed to confirm the generalizability of our conclusions.

2.7 Conclusion and Future Work

In conclusion, we define anonymity from the perspective of early career developers as not sharing identifiable information such as name, picture, location, or professional background, while maintaining a clear separation between one's online and real-life identities and avoiding cross-platform identification. This definition of anonymity is a result of the interview survey with 34 early-career developers after consolidating their individual definitions.

We determined that developers tend to have anonymous accounts on social coding platforms rather than social media and professional platforms. When comparing the developer's behavior in terms of anonymous accounts on social media platforms against social coding platforms and professional platforms we see that developers tend to have anonymous accounts on social coding platforms with a high correlation of 0.65 to avoid associating their activity on the platform with their real-life identity.

We found that the engagement of early-career developers with privacy policies on social coding platforms is generally low. Many developers skim through these policies or

ignore them altogether, often perceiving them as lengthy and complex. However, when privacy policies are presented in a more structured and simplified manner, developers show a greater willingness to engage. This lack of engagement highlights a significant gap between the existence of privacy policies and their practical impact on user behavior.

When privacy policies are restructured using contextual integrity (CI) principles, developers demonstrate an improved understanding of these policies. CI principles, which emphasize the context of information flow and the appropriateness of data sharing, make the policies more relatable and comprehensible to developers. This restructuring leads to better-informed decisions and awareness of privacy implications among developers.

Future research should investigate the definition of anonymity of developers across different career stages. It would be valuable to explore how early career and experienced developers navigate the trade-offs between professional advancement and maintaining their desired level of anonymity. Additionally, understanding the long-term impact of anonymity on developers' career trajectories and online identities can provide insights into how these terms shape the professional landscape and contribute to evolving norms around privacy and anonymity in the tech industry.

Chapter 3

Enigma of Anonymity on Social Coding Platforms

3.1 Introduction

Software developers are increasingly adopting social, collaborative platforms for software development and making a career reputation [60]. These platforms help developers build an online portfolio of their skills by engaging in open source projects and knowledge crowd sourcing [124]. Studies showed that extrinsic motivations, especially regarding career prospects, directly motivate users to contribute, and gamified incentive mechanisms act as a means of realizing that target [63, 17, 115].

Despite the potential advantages of using real identities for professional visibility, many developers choose to operate under pseudonyms and cartoon avatars, hiding their identities in a veil of anonymity [1]. Christopherson [21] defines anonymity as the complete removal of identifying information (such as name) from one's online interactions. Pseudonymity is defined as when a user chooses a name or signifier different from their identity as their identifier for representing their profile and their actions across the site [28]. All these levels of chosen anonymity on social coding platforms are possible as many platforms do not often enforce a strict real name policy [87]. Using pseudonyms and unrealistic avatars (aliases) is not new on social media. Popular social media platforms face an abundance of fake and unreal accounts, some trolling and causing online disturbances for the users [102], which led to research filed for identifying and banning such behaviors [125].

There are a variety of opinions on the impact of developers' online presence on their career development [63]. The debate spills over to online forums, with conflicting advice on real names versus aliases. Several developers have asked a question along the lines of "Should I use my real name on Stack Overflow/GitHub?"[1] on social media. In response to these questions, some believe that using a real name can help "... building a personal brand and establish credibility, while a nickname can provide a level of anonymity and creativity," and others suggest only using real names "... on GitHub and LinkedIn because users consider GitHub and LinkedIn to be part of their resume"[2]. While on Twitter, users recommend "don't use your real traceable name on Stack Overflow because one day you might find that your old posts were wrong and you look stupid in them, but they will stay on the internet forever"[3]. On Stack Exchange, users recommend "to use your own name on all forums, not because they want to push their name for their career but because they want to show that they stand publicly by my opinions"[4]. From the scientific perspective, Ford et al. found that software developers sharing their identity online has advantages since it can increase their visibility online and showcase their talents for future careers [34, 93]. Additionally, research by Sarma et al. suggests that online presence can lead to opportunities for collaboration and knowledge exchange within the developer community [93]. However, some studies justify users not providing their personal information because hate speech and posting offensive comments for personal reasons make the Stack Overflow community "toxic" [20].

This very lack of consensus in both the research world and online forums and motivated by the discourse of research studies and the online forums on the benefits and drawbacks of sharing personal information, we are interested in evaluating the extent to which the developers are anonymous on social coding platforms. We assumed that if a developer's LinkedIn profile can't be retrieved using the information shared on their social coding platforms, then the developer is anonymous. We used LinkedIn, a professional network, for this since it enforces a real-name policy[5] on the platform, and it has been used in research for identifying activities and career paths of software developers [14, 65]. In this context, we answered two research questions:

[1] `https://shorturl.at/XSu71`
[2] `https://shorturl.at/YznSO`
[3] `https://twitter.com/Sc_Meerkat/status/1692893904675209601`
[4] `https://shorturl.at/0HtwL`
[5] `https://www.linkedin.com/legal/user-agreement`

RQ1: To what extent do developers share their identifying information on coding platforms and social networks? *We sample a set of 130,000 developer profiles on Stack Overflow and retrieve their profiles across different platforms to see the availability of the information and the extent to which they share info across different platforms.*

RQ2: To what extent can developers' professional profiles on LinkedIn be retrieved using the information they shared on social coding platforms? *We developed a search and retrieval process to query developers' LinkedIn profiles using the information we retrieved in **RQ1**. This was done to assess whether and to what extent the shared data can contribute to identifying the developers' professional and real identity as relevant to their careers.*

These two questions mark the initial step in comprehending the prevalence of anonymity on social coding platforms. We anticipate that answering these RQs helps in understanding the prevalence of anonymity on social coding platforms. More importantly, they enable us to progress in comprehending the reasons *why developers do not disclose their real identity on these platforms, despite the growing implementation of recognition features*. Additionally, we aim to explore how developers navigate the decision to share their data and how platforms can take action to address any underlying circumstances.

3.2 Related Works

Identifying developers across platform: The work of Fang et al. aimed at helping to provide deeper insights into how developers' actions on one platform are influenced by their roles on other platforms [29]. They employed two heuristics: the first involved gathering Twitter profiles directly linked to GitHub accounts, while the second heuristic focused on crawling personal websites linked from GitHub user-profiles and extracting links to associated Twitter accounts [29]. The authors found that using this heuristic would exhibit relatively high precision in their linkage efforts in merging the identities of the GitHub developers with their Twitter profiles. Vasilescu et al. [112] have tried to address the problem of identifying developers across platforms in order to investigate the switchover rate from R mailing list[6] to StackExchange[7] based on participant overlap.

[6] `https://www.r-project.org/mail.html`
[7] `https://stackexchange.com`

The researchers compiled data from these two distinct sources—StackExchange R subcommunities (StackOverflow and Cross-Validated) and the main R user support mailing list (r-help) [112]. The method the researcher's used inorder to merge users from Stack Exchange and R-mailing lists involved using their email MD5 hashes in order to determine the user turnover rate. They stated that all existing methods to merge identities result in false negatives and false positives [112].

Weise et al. investigated different disambiguation heuristics to see which would work best in determining the ideal identification technique in a given situation [116]. One of the main notable techniques mentioned by them is using Levenshtein similarity for comparison between a pair of full names, first names, last names, and addresses of members of mailing lists with a threshold of 0.93 [116]. The authors used information extracted from different platforms such as JIRA, project websites, and community public keys repository to build a reference baseline, used it to evaluate the disambiguation heuristics, and investigated the time window effect on disambiguation as well. Their findings suggest that the Levenshitien similarity method, which they developed following using the method of Bird et al. [11], performs well in the disambiguation of mailing list users across different platforms.

Anonymity in social coding platforms: With the increasing number of software developers, there is also social coding platforms, such as Stack Overflow [8], which offer a unique space for developers to seek and share information without the necessity of revealing their identity, thanks to the absence of a strict real-name policy [82]. This feature proves valuable for individuals who are cautious about the potential repercussions in their professional or personal lives. However, the downside lies in the challenge of verifying a developer's credibility or expertise due to the lack of identity confirmation, leading to potential issues of misinformation within the community. To address this, Stack Overflow has introduced a user profile page, allowing developers to provide details about their education, work experience, and areas of expertise, thereby enhancing credibility assessment [129]. Despite these measures, it is crucial to acknowledge that not all developers may feel comfortable revealing their identity, emphasizing the need to define and understand the spectrum of anonymity on platforms social coding platforms.

[8] `https://stackoverflow.com`

3.3 Empirical Data and Method

In what follows, we explain the data and protocols for each RQ.

3.3.1 Completeness of developers' profile (RQ1)

Stack Overflow does not enforce a real name policy while having dedicated fields for linking GitHub, Twitter, and personal blogs on developers' profiles. Hence, for the purpose of answering our RQs we started by mining Stack Overflow. We chose to sample a random set of developer profiles to reduce sampling bias [53]. We generated a sample user ID using a `python`-based random number generator and retrieved the user profiles from `Stack Exchange`. We retrieved 130,000 profiles (0.63% of the number of all user profiles on Stack Overflow) using this method. We then removed the deleted and deactivated accounts that returned a NaN status code 404. We also used the PYTHON built-in standard method *isascii*[9] to identify names with non-English characters and exclude those user profiles from our data to simplify the search, as different platforms support non-English characters differently. As a result, we removed 11,754 inactive and 5,808 non-English profiles and hence performed our analysis for **RQ1** and **RQ2** on the remaining 112,438 profiles.

To sign up on Stack Overflow, developers must pick a unique username and provide a screen name. While these are mandatory, users can optionally include details like location, avatar, GitHub, Twitter (now X), and personal website in their profiles, linking GitHub and Twitter explicitly. We used these links to retrieve the developer's accounts on GitHub and Twitter whenever available. We used the GitHub API [39] and Twitter API [122] to retrieve the full user profiles for each user with a linked account. The username is a required text field on Stack Overflow (), GitHub (), and Twitter (). While selecting a screen name is optional on GitHub, it is mandatory on the other platforms. Location, external URLs, and avatar are optional to fill in on all the platforms. It is noteworthy to mention that we did a structural analysis of the screen names to determine if they were pseudonyms or not. We classified the names as pseudonyms if the name was shorter than four characters, contained any numbers, or included any special characters (except hyphens).

[9] `https://docs.python.org/3/library/stdtypes.html`

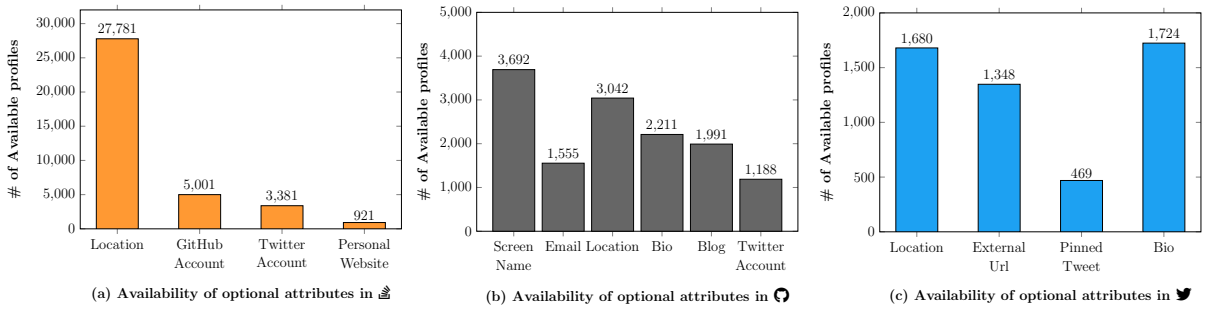


FIGURE 3.1: Availability of "optional" attributes on active sampled Stack Overflow (a), GitHub (b), and Twitter (c) profiles.

3.3.2 Identifying LinkedIn profiles (RQ2)

Given that not all the information the developers provide is exactly the same or available across GitHub and Twitter, we focused on the most commonly shared attributes by developers. Hence, we used the data screen name and location we retrieved from each platform individually and aggregated to evaluate whether we could retrieve developers' LinkedIn profiles. When obtaining a union, there are two cases in which we did not carry out a search on LinkedIn since it was not needed or the query would not return any new results:

1. when the developer did not link their GitHub and/or Twitter accounts in their Stack Overflow profile.

2. when the information on a GitHub and/or Twitter profile provided by the developer is exactly the same with the information provided on Stack Overflow.

Overall, to answer **RQ2**, we carried out eight distinct search queries on LinkedIn for each developer as listed in Table 3.1. We did SS_1 (Search of screen name) and SS_2

TABLE 3.1: Search strings to retrieve LinkedIn profiles in **RQ2**, using Stack Overflow (), GitHub (), and Twitter () data.

Search ID	Query type	# of Dev.
SS_1 ()	"Screen name"	112,438
SS_2 ()	"Screen name" AND "Location"	27,781
SS_1 ()	"Screen name"	3,692
SS_2 ()	"Screen name" AND "Location"	2,949
SS_1 ()	"Screen name"	2,033
SS_2 ()	"Screen name" AND "Location"	1,680
SS_2 (U)	"Screen name" AND "Location"	2,353
SS_2 (U)	"Screen name" AND "Location"	1,431

(Search of screen name and location) using Stack Overflow information. We performed SS_1 (Search of screen name) using GitHub (SS_1 ()). We then performed SS_1 (Search of screen name) using Twitter (SS_1()) attributes. We also carried out SS_2 (Search of screen name and location) using GitHub SS_2() attributes first, then Twitter SS_2()) attributes to identify the LinkedIn profile of the developers. We performed a union of the developers' screen name and location from Stack Overflow & GitHub SS_2(U) to search for developer profiles on LinkedIn. Lastly, we did a union of Stack Overflow & Twitter SS_2(U) screen name and location to perform searches on LinkedIn. The total number of developers along with the search ID and query type we performed these queries on is described in Table 3.1. So we used a metasearch engine known as Searxgn[10] enabling us to have access to 76 online instances, which we used to browse through and perform the search queries in order to avoid having our requests blocked.

3.4 Empirical Results

3.4.1 Completeness of developers' profile (RQ1)

For the 112,438 developers, we gathered all the mandatory and optional fields in their profiles. Figure 3.1 (a) demonstrates the availability of the optional fields while knowing that the mandatory fields are available for all the users. Overall, among the 112,438 developers, 74.55% (83,788) provided none of the optional attributes, while 25.45% (28,650) filled in at least one of the optional fields, and 0.39% (437) filled in all such fields. This suggests a preference for a minimal online footprint, potentially contributing to anonymity.

Among the optional fields, location exhibits the highest availability at 24.7% (27,781), followed by GitHub Link with 4.4% (5,001), Twitter Link with 3% (3,381), and Personal website with the lowest availability at 0.81% (921). It is noteworthy that the avatar field is also an optional attribute that can be provided by the users where Stack Overflow adds a default placeholder avatar (which is a hashed image that signifies the user's IP address) in case the user doesn't provide an image of their choice on the platform. Among the 0.81% of the personal websites provided by the developers, 0.083% (94) were links to their LinkedIn account, while the rest were links to their personal web pages. As mentioned in Section 3.3.1, we checked whether the screen names of the Stack Overflow developers

[10] `https://docs.searxng.org`

could be classified as pseudonyms. We identified 55,107 (49.1%) developers who use pseudonyms on Stack Overflow from a total of 112,438.

A total of 5,001 developers have self-provided their GitHub accounts, and 3,381 developers have self-provided their Twitter accounts in their Stack Overflow profiles. Among them, 2.3% (2,583 developers) linked both GitHub and Twitter accounts, of which 1,406 were active. Out of these 5,001 GitHub accounts, 4,156 are active. Among these developers (See Figure 3.1 (b)), 88.8% provided a Screen name, 73.1% shared their location, 37.4% shared email, 47.9% provided a bio description, 53.2% provided their blog accounts, and 28.5% also linked their Twitter account on GitHub. Similar to Stack Overflow, on GitHub, as well the after screen name, the most provided optional attribute is location, with 73.1%(3,692) profiles having shared it. Out of the 3,381 linked Twitter accounts, 2,033 were active. On Twitter, choosing a screen name is mandatory; hence, we gathered it for all the developers. Among the developers with active accounts (See Figure 3.1 (c)), 82.8% of the developers have shared their location, 66.3% shared their External URL, 23.1% had Pinned tweets, and 84.8% also had a bio description. In similarity with Stack Overflow and GitHub, on Twitter, after the bio description, the most available optional attribute is location. Analyzing these linked accounts revealed similar trends: location and bio descriptions were commonly shared, while email addresses and personal websites remained less frequent.

RQ1 Main Result: In our sample, 2.3% of developers linked both accounts, with 1,406 being active. 73.1% of GitHub users and 82.8% of Twitter users shared location, while 88.8% of the GitHub users shared screen name. Profile completion is low across platform where it is 0.39%, 2.21%, 16.7% on Stack Overflow, GitHub and Twitter respectively. This suggests that there is a clear preference for partial self-disclosure among developers on social coding platforms.

3.4.2 Identifying LinkedIn profiles (RQ2)

We performed a total of eight different queries in order to answer **RQ2** fully (see Table 3.1). We performed screen name and location queries on LinkedIn using only Stack Overflow attributes. We also conducted queries using specific search spaces (SS_1 and SS_2) on

GitHub and Twitter to retrieve developers' LinkedIn profiles. In addition, we also conducted aggregated search queries of the developers as mentioned in Section 3.3.2. The result of these eight search queries is described in Table 3.2. The analysis categorized developers based on the number of LinkedIn profiles retrieved, where the first column provides the type of the query, while the rest of the columns show the number of developers where we found zero LinkedIn profiles, one LinkedIn profile, or more than one LinkedIn profile, respectively.

We have shown the developers we were uniquely able to identify on LinkedIn exactly one account in the upset plot (See Figure 3.2). When comparing the group of developers, we determined exactly one LinkedIn profile when using SS_2 of Stack Overflow to the developers who had one LinkedIn profile when using SS_2 on the union of Stack Overflow and GitHub attributes. We can see that there are 1,594 (See Figure 3.2) and 95 developers, respectively (See Figure 3.2). When comparing these two groups, we found only one common developer out of the 95 developers to the 1,594 developers, which also had exactly one LinkedIn profile when using SS_2. Our findings indicate a noteworthy trend: the inclusion of supplementary information, specifically GitHub data in union with Stack Overflow data, appears to diminish rather than enhance developer findability. This is because developers do not consistently share information across different platforms and the use of this inconsistent information introduces noise. This observation raises questions about the nature of the data and the impact of combining information from different sources on the findability of developers.

Similarly, when comparing the developers, we determined exactly one LinkedIn profile, using SS_2 of Stack Overflow alone to the developers who had one LinkedIn profile

TABLE 3.2: Identified users on LinkedIn (in) with Stack Overflow (), GitHub (), Twitter() and aggregated attributes

Search ID	0 in profiles	1 in profiles	>=2 in profiles
SS_1()	45,345 | 40.3%	5,567 | 4.9%	61,526 | 54.8%
SS_2()	15,716 | 56.6%	1,594 | 5.7%	10,471 | 37.7%
SS_1()	283 | 7.7%	323 | 8.7%	3,086 | 83.6%
SS_2()	209 | 7.1%	280 | 9.5%	2,460 | 83.4%
SS_1()	351 | 17.2%	191 | 9.4%	1,491 | 73.4%
SS_2()	292 | 17.4%	187 | 11.1%	1,201 | 71.5%
SS_2(U)	1,760 | 74.8%	95 | 4.0%	498 | 21.2%
SS_2(U)	1,052 | 73.6%	51 | 3.6%	328 | 22.8%

using SS_2 on the combination of Stack Overflow and Twitter attributes there are 1,594 (See Figure 3.2) and 51 developers respectively (See Figure 3.2). When comparing these two groups, we found that there are only two common developers out of the 51 developers to the 1,594 developers who had exactly one LinkedIn account using SS_2. This further strengthens our point, which is that the additional data seemingly introduces a heightened level of anonymity, which is counterintuitive.

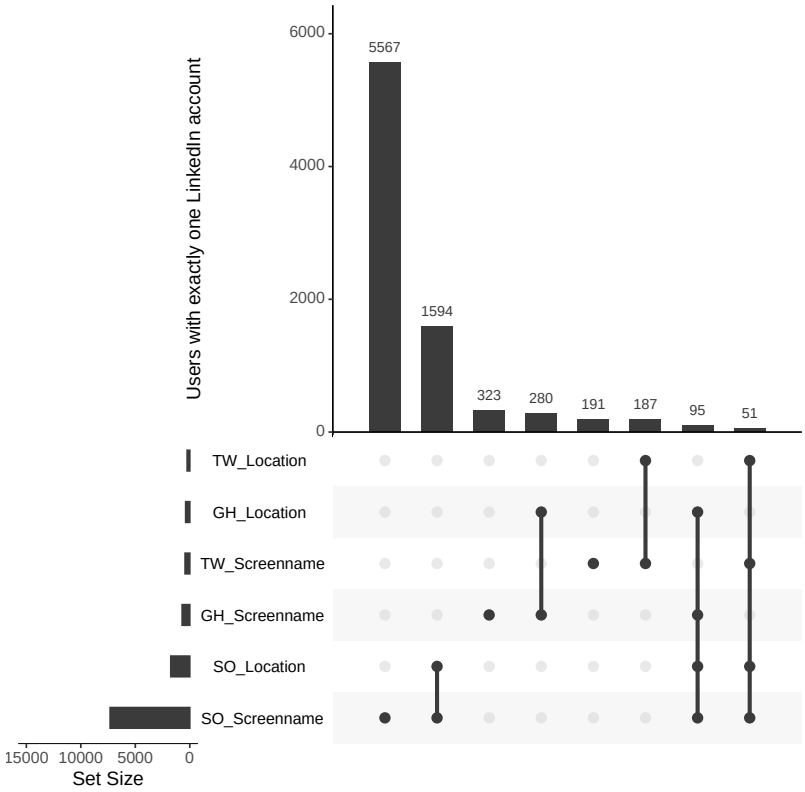


FIGURE 3.2: Findablity of unique profiles on LinkedIn

RQ2 Main Result: Our findings suggest that using Stack Overflow location alongside screen name for the search query on LinkedIn helps narrow down developers' profiles. However, the combination of cross-platform attributes from both Stack Overflow and GitHub, as well as Stack Overflow and Twitter, introduces noise or conflicting information, making it more difficult to narrow down the developer's LinkedIn profiles.

3.5 Discussions and Emerging Research

Reflecting on the results of our **RQ1** and **RQ2**, we see several potential future research areas.

3.5.1 Is developers' anonymity intentional or not?

While we demonstrated that for 93.5% of the sampled developers, we couldn't identify their LinkedIn profile, a comprehensive diagnostic study is required to determine whether this anonymity is intentional, aimed at concealing identity, or accidental, arising from a lack of effort in providing data, for instance. We performed a preliminary qualitative study with the developers to determine if this anonymity was intentional (see Table 3.3).

We designed a very short survey and advertised it on our networks for a convenient sampling. 117 software developers (97 Males (83.0%), 19 Females (16,2%), and one other (0.9%)) responded to the survey. Among them, 114 (97.4%) of 117 developers had GitHub accounts. We also asked if they had Stack Overflow accounts, of which 19 (16.2%) of them had an account, and 11 (9.4%) of them asked questions on Stack Overflow. Nine of the developers stated that they are anonymous on the platform and no one can identify them since they use pseudonyms, while eight (6.8%) of them didn't intentionally hide their identity, and two (1.7%) were identifiable since they used their real name. It is worth noting that all the participants who were not male-identifying were anonymous or never had an account.

Surveyed developers demonstrated a significantly greater number of accounts on GitHub than on Stack Overflow. This highlights the need for additional exploration into the importance of various activities and their impact on developers' engagement. Additionally, an analysis of the requirements and significance of logged-in profiles for accessing platform features is warranted.

3.5.2 Is anonymity beneficial for developers' careers or not?

In section 3.1, we presented existing research and online forums that brought forward conflicting views on which approach developers should take. While some advocate for

TABLE 3.3: Results of surveying 117 software developers to determine if they are identifiable or anonymous on Stack Overflow

Survey Question	%
Q1: Do you have a GitHub account?	
Yes	97.4%
No (not applicable)	2.6%
Q2: Do you have a Stack Overflow account?	
Yes	16.2%
No	83.8%
Q3: Do you post questions on Stack Overflow?	
Yes	9.4%
No	90.6%
Q4: Are you anonymous on Stack Overflow?	
I am anonymous. No one can identify who I am.	7.7%
Not sure, but I did not intent on hiding my identity.	6.8%
I am identifiable and I am using my real name.	1.7%
Not Applicable	83.8%
Q5. What is your self-identified gender?	
Male	83.9%
Female	16.2%
Other	0.9%

real-name usage to enhance visibility and recognition, others highlight potential downsides like repercussions from past mistakes. We have also presented that research done by Lu et al. [63], the recognition of the social coding platforms though matters in the sense that it is beneficial for their careers [63]. Our findings showed that on average for 94.7% of the sampled developers from Stack Overflow, one cannot retrieve their definite LinkedIn profile. Among developers who linked their GitHub and/or Twitter, we still could not identify their definite LinkedIn Profiles on average for 92.3% of them (see Table 3.2). This indefinite identification of LinkedIn profiles points to the potential anonymity of overall 93.5% developers in our sample. This is despite the potential benefit of online recognition and reputation that activities on these platforms can provide to the developers.

Our research doesn't provide a definitive answer as to which approach is better. However, our vision with this paper is to motivate further investigation because the low profile completion rate suggests a potential disconnect between the perceived career benefits and developer practices. We believe that qualitative research with developers and recruiters can shed further light on anonymity's potential benefits and drawbacks on these platforms. Future studies should take into account diverse career stages and various development activities (such as questioning versus code contributions) and consider the perspectives of different marginalized groups [34].

3.5.3 Why do developers feel the need to be anonymous?

Understanding why developers choose to be anonymous is a complex issue that needs further exploration. While our current study highlights differences in how much information developers share on different platforms, it doesn't explain the reasons behind their preference for anonymity. The results from **RQ1** showed that there are differences in how much information developers share on different platforms. These variations could be because of the rules each platform follows, including their data-sharing practices. Some developers may choose to stay anonymous on purpose, but the low profile completion rate could also be influenced by how the platform is designed. For example, the Stack Overflow privacy policy states that "*Users should be aware that such content, once public, cannot often be removed from public view even if it is removed from the Stack Overflow network and our servers*". Our findings suggest a significant portion of developers choose partial self-disclosure on Stack Overflow (see Section 3.4 and Table 3.3). This raises questions about the role of platform policies in shaping user behavior.

To uncover this, future research could use interviews or surveys to directly ask developers about their choices. It's essential to explore how individual decisions, platform rules, and potential effects on career development interact. Research should also consider different career stages, types of development activities, and the perspectives of various groups to get a more detailed understanding. This approach can contribute to creating more inclusive and supportive platforms for developers.

Our analysis revealed that a significant number of information flows described within the privacy policies of Stack Overflow and GitHub were incomplete. These incomplete information flows within privacy policies can make it difficult for users to understand exactly what data is collected, how it's used, and with whom it's shared. This lack of clarity could contribute to user hesitation in completing profiles, leading to a preference for partial self-disclosure or even unintentional anonymity. These findings highlight the importance of clear and comprehensive privacy policies on social coding platforms. Platforms should strive to communicate data practices in a way that is easily understandable for users. Additionally, future research can explore deeper into the interplay between platform design, privacy policies, and user behavior regarding identity disclosure on social coding platforms.

3.5.4 Does consistency of the information developers share play a part in anonymity?

To assess the consistency of information across developers' accounts, we focused on the developers' linked accounts (i.e., the GitHub and/or Twitter accounts linked on their Stack Overflow profile). Among these profiles, our attention was directed to data that was commonly available in all profiles, including username, screen name, location, and avatar images. To answer RQ2 and assess the consistency of information between each two accounts, we performed pairwise comparisons. In what follows, we explain the steps taken for text and image comparisons.

Comparing textual attributes:

For the text attributes (screen name, username, and location), we employed both exact string matching of two text strings and substring matching. The exact string matching function determines whether the text of two strings (e.g., screen name on Stack Overflow and screen name on Twitter) matches character by character, disregarding special characters such as backslashes and underscores without being case-sensitive. For partial string matching, we employed substring matching. In this process, we determined whether the characters of the substring appear consecutively and in the same order within the larger string. It is important to note that contiguous placement of characters is not required, but their sequential order must be preserved.

Algorithm 1 shows both the process of exact and partial matching. We acknowledge there could be special cases for the location attribute where the developers could use acronyms instead of the full the names of the countries when sharing location. Hence, we use the PYTHON built-in package called pycountry [11] to convert the acronyms to the corresponding full names of the countries. We use this step before we perform substring matching, as outlined in Algorithm 1. We perform this preprocessing step so that the accuracy and completeness of our analysis increases especially when dealing with variations in the representation of location information in our dataset.

[11] `https://pypi.org/project/pycountry/`

Algorithm 1 Evaluating text attributes for consistency

```
function ExactMatching(A, B)
    if A = B then
        return True
    else
        return False
    end if
end function

function SubstringMatching(A, B)
    for i ← 0 to length(A) − length(B) do
        if A[i : i + length(B)] = B then
            return True
        end if
    end for
    return False
end function
```

Comparing images (avatars) across platforms:

Avatars are digital representations of users that may be abstract, cartoonish, or human-like [83]. We analyzed avatars to determine whether developers used the exact same photo across different platforms and whether the subject featured in one avatar corresponds to the subject in alternative avatars on other platforms (i.e., whether the same person is depicted in different photos). To analyze the consistency of avatars across different platforms, we used three different methods namely Multistructural convolutional network (MTCNN) [127], Structural similarity (SSIM) [101], and DeepFace [97].

Multistructural Convolutional Network (MTCNN) [18]: is a face detector model for KERAS, which was developed from scratch based on Facebook's reference implementation of MTCNN which is known as Facenet [96]. MTCNN retrieves key points such as the nose, left of the mouth, right of the mouth, and left eye from an image, then removes the parts it deems that are not part of the face and gives an output facial dimension of the image for detection [127]. The benchmark dataset for MTCNN uses frontal face images for detection, which achieves an average precision of 0.76 [18]. Since MTCNN is only a face detection model that identifies and locates the face, we used a state-of-the-art PYTHON face recognition [38, 126] to compare the returned facial dimensions by MTCNN to calculate the similarity score between a pair of images. This model has an accuracy of 99.38% on the Labeled Faces in the Wild [108] benchmark. The similarity scores returned

by this model range between 0 and 1, where the higher values mean the pair of images has more similarity.

Structural Similarity Index (SSIM) [101]: SSIM quantifies the likeness between two images by assessing their luminance, contrast, and structural components. Unlike MTCNN, which focuses on detecting specific key points for face detection, SSIM evaluates the overall content of the entire image for comparison [114], which aids in identifying different versions of the same image. SSIM effectively notices small changes in images, which is valuable in situations where maintaining the visual quality is essential. This is especially useful when dealing with tasks like image compression, such as when uploading a photo as a profile picture on a platform. Higher SSIM values indicate a greater degree of similarity, ranging from 0 (no similarity) to 1 (perfect similarity). Conversely, a perfect dissimilarity is denoted by an SSIM value of -1.

DeepFace [85]: DeepFace[12] is an open-source facial recognition library developed by Facebook's AI research team in Python. DeepFace uses Convolutional Neural Networks (CNNs) to perform face verification and recognition tasks. DeepFace has high precision in detecting faces in different lighting, poses, and facial expressions. This model provides a hybrid face recognition framework wrapping state-of-the-art VGG-Face [86], Google FaceNet [96], OpenFace [3], Facebook DeepFace [104], DeepID [103], ArcFace [25], and D-lib [126]. In particular, we used the VGG-Face model [86] for our purposes because it has high accuracy in both the Labeled Faces in the Wild benchmark [108] and the YouTube Faces benchmark [120], which are 98.78% and 97.4%, respectively. The similarity scores returned by the DeepFace model range between zero and one, where the higher values show that the pair of images has more similarity.

Image comparison ground truth setup

We used a set of users who had originally self-identified their LinkedIn profiles on Stack Overflow, which are 94 users in number. We refer to this dataset as the ground truth dataset. We retrieved the images of these users from Stack Overflow, GitHub, Twitter, and LinkedIn. Then, we formed a dataset of 248 rows from 94 users where they could have images from different platforms (this includes Stack Overflow, GitHub, Twitter, and

[12] `https://github.com/serengil/deepface`

LinkedIn). We ran the algorithms MTCNN [18], SSIM [101], and DeepFace[13] by using pairwise comparison of the images of these users across the platform. Then, we used the first condition: a threshold of 0.5 for DeepFace and a threshold of zero for SSIM to consider similar images. We then used our second condition to consider the pairwise images of the users if it failed the first condition. The second condition we had in place was to see if the images had a similarity above 0.3 for MTCNN to consider the users' images as the same person. This resulted in the result of 54% of the users having satisfied the conditions. Then, we expanded our analysis on all the images of the users from Stack Overflow, GitHub, and Twitter, which can be seen in Table 3.4.

The check we ran to see if we had the exact same images was to see if at least one of the similarity models, which are MTCNN, SSIM, and DeepFace, had a value of one in the ground truth dataset. This amounted to about 10% of the developers from the ground truth data set. We ran this analysis on all the images of the users from Stack Overflow, GitHub, and Twitter, which can be seen in Table 3.4.

We employed these three methods in complement with each other to detect the similarity between a pair of images that a developer has utilized across various platforms, aiming to ascertain if the subject in the three images is the same. This comprehensive approach ensures the strength and flexibility of our image similarity assessment, allowing it to handle differences in image content, orientation, and structure. Whether users upload images with standard front-facing views or opt for more diverse compositions, our method is capable of identifying similarities even between cartoonish avatars or avatars that have undergone decompression and alterations or have significantly different resolutions. The average correlation value between MTCNN and SSIM is 0.37 for the pair of images across platforms, which are Stack Overflow, Twitter, and GitHub. The average correlation value between SSIM and DeepFace is 0.78, and the average correlation value between MTCNN and DeepFace is 0.19.

Consistency of textual and image attributes across platforms

When looking across Stack Overflow, GitHub, and Twitter, the common attributes across the platforms for these users were Screen name, Username, Location, and Avatar. For each user on Stack Overflow, we investigated if and to what extent these accounts were

[13] `https://github.com/serengil/deepface`

linked and had consistencies for each information field. We present the results of this analysis for both the text and image attributes in Table 3.4.

According to the results in Table 3.4, developers with both Stack Overflow and GitHub accounts that have consistently shared the exact same value when sharing their username are 27.9%, whereas the 72.1% have inconsistently shared their username (refer to Exact-Matching Algorithm 1). The developers that consistently share the exact same screen name make up 46.4%, and ones that don't amount to 53.6% when the screen name attribute (refer to ExactMatching Algorithm 1). The developers that have shared at least a partially consistent (similar) screen name amounts to 69.2%, whereas the developers that haven't are 30.8% (refer to SubstringMatching Algorithm 1). Regarding the consistency of the location attribute for the developers 37.9% have consistently used the same exact location, while 62.1% haven't (refer to ExactMatching Algorithm 1). In addition, 57.1% of these developers shared at least a partially consistent (similar) value for location while 42.9% have not done so (refer to SubstringMatching Algorithm 1). In terms of avatar usage, 17.4% of developers have used the same exact image for both their Stack Overflow and GitHub accounts, whereas 82.6% haven't. Our avatar comparison algorithm constructed from MTCNN [18], SSIM [101], and DeepFace [85] has identified 76.1% of developers as the same person, while 23.9% of them aren't.

The developers with both Stack Overflow and Twitter accounts (refer to Table 3.4), that have consistently shared the exact same username are 19.9%, whereas 80.1% have not used the same exact value (refer to ExactMatching Algorithm 1). In terms of these developers' consistency with the Screen name attribute (exact match), ones that have used the exact same value make up 47.7%, while 52.3% haven't done so (refer to ExactMatching Algorithm 1). The developers that have shared at least a partially consistent (similar) screen name amounts to, 74.1%, whereas 25.9% haven't used partially similar screen names (refer to SubstringMatching Algorithm 1). For the case of location, developers that have shared the same exact value amount to 31.1%, while 68.9% of the users haven't (refer to ExactMatching Algorithm 1). Further, 56.4%, of these developers have used at least a partially similar location value while 43.6% have not done so (refer to Substring-Matching Algorithm 1). In terms of avatar usage, 0.3% of developers have used the same exact image for both their Stack Overflow and Twitter accounts, while 99.7% haven't. Our avatar comparison algorithm which we constructed from MTCNN [18], SSIM [101],

TABLE 3.4: Consistency of Text & Image Attributes on

Attribute	Vs	Vs	Vs
Username	27.9%	19.9%	61.9%
Screen Name *(exact)*	46.4%	47.7%	69.3%
Screen Name *(substring)*	69.2%	74.1%	84.8%
Location *(exact)*	37.9%	31.1%	43.5%
Location *(substring)*	57.1%	56.4%	62.3%
Exact Same Image	17.4%	0.3%	0.2%
Same Person	76.1%	64.3%	62.5%

and DeepFace [85] has detected 64.3% of developers as the same person, while 35.7% of users aren't.

The developers with both GitHub and Twitter accounts, that have consistently shared the exact same username are 61.9%, while 38.1% of them have not used the exact same values (refer to ExactMatching Algorithm 1). In terms of these developers' consistency with the Screen name attribute (exact match), ones that have used the exact same value make up 69.3%, whereas 30.7% have't (refer to ExactMatching Algorithm 1). The developers that have shared at least a partially consistent (similar) screen name amounts to 84.4% while 15.2% of them have not done so (refer to SubstringMatching Algorithm 1). Regarding the consistency of the location attribute for the developers 43.5% have used the same exact value 56.5% haven't. In addition, 62.3% of these developers shared at least a partially consistent (similar) value for location while 37.7% haven't In terms of avatar usage, 0.2% of developers have used the same exact image for both their GitHub and Twitter accounts, while 99.8% haven't. Our avatar comparison algorithm which we constructed from MTCNN [18], SSIM [101], and DeepFace [85] has detected 62.5% of developers as the same person, while 37.5% of users aren't.

In addition, we performed a comparison of the avatars the developers shared on Stack Overflow and LinkedIn to see if developers shared their pictures. From the 5,567 developers we identified in Section we found that 1,047 of the developers used non-default profile pictures on both Stack Overflow and LinkedIn. When we ran our avatar comparison which we constructed from MTCNN [18], SSIM [101], Deepface [85] detected that 38.7% of the developers as the same person and 10.5% of the developers using that same exact image on Stack Overflow and LinkedIn (See Table 3.5).

3.6 Threats to Validity

In this section of the paper, we will describe the threats to the validity of the research as described by Wohlin et al. [119].

Our analysis relied on self-identified developer accounts on GitHub and Twitter, linked to Stack Overflow data. While this approach yielded results for 4.4% of GitHub users and 3% of Twitter users, it excludes a significant portion of the original Stack Overflow sample (approximately 96%). This limitation could potentially impact the generalizability of our conclusions. Instead of relying solely on self-identified accounts, we could have actively queried user accounts on GitHub and Twitter using their Stack Overflow profile attributes directly from the platforms. This could have potentially identified more users on these platforms beyond the self-identified ones, leading to a larger sample size. Future research could explore these options to expand the sample and enhance the generalizability of our findings.

Our research questions rely on profile attributes like screen name and location. These were chosen because they were the most available attributes across Stack Overflow, GitHub, and Twitter. However, this choice necessitates acknowledging potential limitations. For instance, screen names might not always reflect a developer's real name, and location data may not pinpoint their specific work location. Additional attributes like username and profile image could be explored to enhance the identification of developers, particularly on LinkedIn. Usernames on LinkedIn are typically tied to real names, offering a potentially more reliable identifier. It's important to acknowledge that even with these additional attributes, challenges might persist. Since usernames can still be pseudonyms, and profile images might not always be present or accurate. Future studies could explore the effectiveness and precision of incorporating more data for the analysis.

TABLE 3.5: Comparison of Stack Overflow and LinkedIn users Image Attributes

Exact Same Image	10.5%
Same Person	38.7%

3.7 Conclusion

We examined developers' profiles on Stack Overflow and their impact on cross-platform identification, particularly on LinkedIn. We discovered varying degrees of data sharing across different platforms. Notably, for 93.5% of the sampled developers, identifying their professional profiles on LinkedIn proved challenging. The prevalent use of pseudonyms on these platforms raises questions about why developers opt for this approach, potentially linking to issues of gender representation and privacy concerns. Additionally, we shared the outcomes of a brief survey with developers, revealing their intentional decision not to disclose identifying information.

Future work should focus on refining how we identify developers across platforms, conducting deeper analyses into gender dynamics, and addressing ethical considerations around developer privacy. Our study revealed increased ambiguity when combining data from different sources, highlighting the intricate nature of online identity management. Our results can help future research refine identification methods, explore gender-related aspects, and explore the ethical dimensions of user privacy in the evolving landscape of developer communities. This work contributes to our understanding of how developers manage their online identities.

3.8 Future Work

Our goal for future work is to expand the data size from over 130,000 users, which will enable an even more comprehensive study. A more comprehensive analysis of the anonymity levels within the developer community will be possible with the larger dataset. The objective of this extension is to incorporate a wider range of experiences, levels of skill, and patterns of collaboration among developers, thereby strengthening the validity, generiziablity, and applicability of our results. We believe that a larger dataset will provide more comprehensive insights into the dynamics of anonymity in developer communities and promote a more comprehensive knowledge of individual notions of anonymity and identification practices.

Chapter 4

Conclusion

Anonymity is an important aspect of the digital world, especially on social coding platforms where developers often navigate the balance between maintaining privacy and building a professional reputation [115]. This thesis aimed to explore how developers define and perceive anonymity on these platforms, examine their engagement with privacy policies and assess the impact of presenting these policies in a different format. Additionally, this thesis aimed to determine the extent to which developers share identifying information on coding platforms and how effectively their professional profiles can be retrieved. Given that this thesis is manuscript-based, consisting of an interview study and a mining study, the conclusions are presented accordingly.

In the interview study of 34 early-career developers, we found that anonymity is defined as the state of not sharing identifiable information such as name, picture, location, or professional background, while maintaining a clear separation between their online and real-life identities and avoiding cross-platform identification. When comparing developers' behavior regarding anonymous accounts on social media platforms versus social coding and professional platforms, we observed a strong tendency among developers to maintain anonymous accounts on social coding platforms, with a high correlation of 0.65, to avoid associating their platform activities with their real-life identities.

We discovered that, on average, none of the early-career developers in the study read the privacy policies of Stack Overflow and GitHub. The primary reasons cited by the developers were the lengthiness of the documents (71%), trust in the platforms (12%), and the necessity of using the platform (9%). However, when the privacy policies were restructured using the Contextual Integrity framework, there was a significant increase

in developers' willingness to read them. We observed a 22% increase, on average, in the likelihood that developers would read the restructured privacy policies of Stack Overflow and GitHub compared to their original formats.

In the mining study of 130,000 Stack Overflow developers, we examined the completion of their user profiles and the identification of their LinkedIn profiles through search queries. Our analysis revealed that 2.3% of developers linked both their Stack Overflow and GitHub accounts, with 1,406 being active users. Among these, 73.1% of GitHub users and 82.8% of Twitter users shared their location, while 88.8% of GitHub users disclosed their screen names. However, profile completion rates were low across platforms, at 0.39% on Stack Overflow, 2.21% on GitHub, and 16.7% on Twitter. This indicates a tendency among developers to provide partial self-disclosure on social coding platforms.

Regarding the identification of LinkedIn profiles, our findings suggest that using the Stack Overflow location in combination with the screen name for search queries on LinkedIn effectively narrows down developer profiles. However, using a combination of cross-platform attributes from Stack Overflow and GitHub, or Stack Overflow and Twitter, introduces noise and conflicting information, making it more challenging to accurately identify the LinkedIn profiles of the developers. By addressing our six research questions with their objectives, we have contributed to the ongoing discourse on the impact of anonymity in the digital age, particularly in professional and collaborative platforms.

4.1 Future Work

Our study has several limitations that should be addressed in future research. One significant limitation is the challenge of generalizing findings due to the complexity of online identity management across various platforms. Additionally, the analysis of different attributes for building a comprehensive model of developer identity remains a challenge. The ambiguity and noise introduced when combining data from multiple sources emphasizes the need for better methods. Future work should focus on refining identification methods across platforms, especially in dealing with the complexities of linking profiles with varying degrees of anonymity. This includes addressing the ethical considerations

surrounding developer privacy and the potential risks associated with cross-platform identification.

Expanding on our qualitative findings, future research could incorporate broader surveys that go beyond early-career developers to include intermediate and experienced professionals. Such surveys would provide a more comprehensive understanding of how different levels of experience influence perceptions of anonymity and engagement with privacy policies. This could help in designing tools and strategies that better support developers at various stages of their careers. Moreover, the development of tools that leverage the Contextual Integrity framework could be a significant step forward. These tools could enhance the readability and relevance of privacy policies, making them more accessible and understandable for developers. Integrating Contextual Integrity principles into privacy policy design could improve developers' engagement and comprehension, thereby fostering a more balanced approach to privacy and professional visibility.

Another area for future research is interviewing employers to understand their perspectives on anonymity and privacy in the context of hiring and professional networking. Insights from employers could provide valuable information on how anonymity impacts developers' career prospects and how privacy concerns are managed from an employer's standpoint. In addition, exploring the ethical dimensions of privacy in developer communities is crucial. Future research should investigate how developers' choices regarding anonymity and pseudonymity (nickname) relate to broader issues of privacy, representation, and inclusion, particularly in the context of gender dynamics and other demographic factors. In summary, future work should aim to build upon our findings by developing more sophisticated tools for managing online identities, improving privacy policies through Contextual Integrity, and exploring the perspectives of various stakeholders, including employers. These efforts will contribute to a deeper understanding of how developers navigate their online presence and help create supportive environments that keep them informed and allow professional growth in the digital age.

Bibliography

[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. “Privacy and human behavior in the age of information”. en. In: *Science (New York, N.Y.)* 347.6221 (2015), 509–514. ISSN: 0036-8075. DOI: 10.1126/science.aaa1465. URL: http://dx.doi.org/10.1126/science.aaa1465.

[2] Muhammad Ovais Ahmad, Iftikhar Ahmad, and Fawad Qayum. “Early career software developers and work preferences in software engineering”. In: *Journal of Software: Evolution and Process* 36 (Oct. 2022). DOI: 10.1002/smr.2513.

[3] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. *OpenFace: A general-purpose face recognition library with mobile applications*. Tech. rep. CMU-CS-16-118, CMU School of Computer Science, 2016.

[4] Areti Ampatzoglou et al. “Measuring and mitigating threats to validity in empirical software engineering research”. In: *Information and Software Technology* 101 (2019), pp. 35–57.

[5] Denise Anthony, Sean W. Smith, and Timothy Williamson. “Reputation and reliability in collective goods: The case of the online encyclopedia Wikipedia”. en. In: *Rationality and society* 21.3 (2009), 283–306. ISSN: 1043-4631. DOI: 10.1177/1043463109336804. URL: http://dx.doi.org/10.1177/1043463109336804.

[6] Ofer Arazy et al. “Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict”. In: *J. Manage. Inf. Syst.* 27.4 (2011), 71–98. ISSN: 0742-1222. DOI: 10.2753/MIS0742-1222270403. URL: https://doi.org/10.2753/MIS0742-1222270403.

[7] Ali Sajedi Badashian et al. “Involvement, contribution and influence in GitHub and stack overflow”. In: *Proceedings of 24th Annual International Conference on Computer Science and Software Engineering*. USA: IBM Corp., 2014, 19–33.

[8] Vinayshekhar Bannihatti Kumar et al. “Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text”. In: *Proceedings of The Web Conference 2020*. WWW ’20. Taipei, Taiwan: Association for Computing Machinery, 2020, 1943–1954. ISBN: 9781450370233. DOI: 10.1145/3366423.3380262. URL: https://doi.org/10.1145/3366423.3380262.

[9] Anton Barua, Stephen W. Thomas, and Ahmed E. Hassan. “What are developers talking about? An analysis of topics and trends in Stack Overflow”. en. In: *Empirical Software Engineer* 19.3 (2014), 619–654. ISSN: 1382-3256. DOI: 10.1007/s10664-012-9231-y. URL: http://dx.doi.org/10.1007/s10664-012-9231-y.

[10] Qerkin Berisha and Aleksandra Klich. “Remote work regulations in the EU, Poland and Kosovo with some considerations from the perspective of the GDPR”. en. In: *Białostockie Studia Prawnicze* 29.2 (2024), 99–117. ISSN: 2719-9452. DOI: 10.15290/bsp.2024.29.02.08. URL: http://dx.doi.org/10.15290/bsp.2024.29.02.08.

[11] Christian Bird et al. “Mining email social networks”. In: *MSR*. New York, NY, USA: ACM, 2006. ISBN: 9781595933973.

[12] Sue Lacey Bryant, Andrea Forte, and Amy Bruckman. “Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia”. In: *Proceedings of the 2005 ACM International Conference on Supporting Group Work* (2005). URL: https://api.semanticscholar.org/CorpusID:221349.

[13] Julio C. Caiza et al. “Reusable Elements for the Systematic Design of Privacy-Friendly Information Systems: A Mapping Study”. In: *IEEE Access* 7 (2019), pp. 66512–66535. DOI: 10.1109/ACCESS.2019.2918003.

[14] Thomas Louis Case et al. “A linkedin analysis of career paths of information systems alumni”. In: *The Journal of the Southern Association for Information Systems* 1.1 (2013), pp. 1–13.

[15] Ann Cavoukian. "Privacy by Design: The 7 Foundational Principles. Information and privacy commissioner of Ontario_2009". In: *Canada* 5 (2009), pp. 405–413. URL: https://api.semanticscholar.org/CorpusID:145599899.

[16] Ann Cavoukian, Scott Taylor, and Martin Abrams. "Privacy by Design: essential for organizational accountability and strong business practices". In: *Identity in the Information Society* 3 (2010), pp. 405–413. URL: https://api.semanticscholar.org/CorpusID:145599899.

[17] Huseyin Cavusoglu, Zhuolun Li, and Ke-Wei Huang. "Can gamification motivate voluntary contributions?: The case of StackOverflow QA community". In: *Conference Companion on Computer Supported Cooperative Work Social Computing*. New York, NY, USA: ACM, 2015. ISBN: 9781450329460.

[18] Iván de Paz Centeno. *MultiStructural Similarity Convolutional Network*. en. 2016. URL: https://github.com/ipazc/mtcnn.

[19] Luke Younghoon Chang, Siew Fan Wong, and Hwansoo Lee. "UNDERSTANDING PERCEIVED PRIVACY: A PRIVACY BOUNDARY MANAGEMENT MODEL". In: July 2015.

[20] Jithin Cheriyan, Bastin Tony Roy Savarimuthu, and Stephen Cranefield. "Norm violation in online communities – A study of Stack Overflow comments". 2020. URL: http://arxiv.org/abs/2004.05589.

[21] Kimberly M. Christopherson. "The positive and negative implications of anonymity in Internet social interactions: "On the Internet, Nobody Knows You're a Dog"". en. In: *Computers in human behavior* 23.6 (2007), 3038–3056. ISSN: 0747-5632. DOI: 10.1016/j.chb.2006.09.001. URL: http://dx.doi.org/10.1016/j.chb.2006.09.001.

[22] Kimberle Crenshaw. "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color". In: *Stanford Law Review* 43.6 (1991), pp. 1241–1299. ISSN: 00389765. URL: http://www.jstor.org/stable/1229039 (visited on 08/03/2024).

[23] Laura Dabbish et al. “Social coding in GitHub: transparency and collaboration in an open software repository”. In: *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM. 2012, pp. 1277–1286.

[24] Martin Degeling et al. “We Value Your Privacy ... Now Take Some Cookies”. In: *Informatik Spektrum* 42 (2018), pp. 345 –346. URL: https://api.semanticscholar.org/CorpusID:52011984.

[25] Jiankang Deng et al. “ArcFace: Additive angular margin loss for deep face recognition”. In: *CVPR*. IEEE, 2019. ISBN: 9781728132938.

[26] Shuili Du and Chunyan Xie. “Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities”. In: *Journal of Business Research* 129 (Aug. 2020). DOI: 10.1016/j.jbusres.2020.08.024.

[27] Paul W. Eastwick and Wendi L. Gardner. “Is it a game? Evidence for social influence in the virtual world”. In: *Social Influence* 4 (2009), pp. 18 –32. URL: https://api.semanticscholar.org/CorpusID:54711350.

[28] EV Charging Energy. *Anonymity and pseudonymity protecting user privacy in online communication*. en. 2023. URL: https://utilitiesone.com/anonymity-and-pseudonymity-protecting-user-privacy-in-online-communication.

[29] Hongbo Fang et al. “Need for Tweet: How Open Source Developers Talk About Their GitHub Work on Twitter”. In: *MSR*. 2020, pp. 322–326. DOI: 10.1145/3379597.3387466.

[30] Samer Faraj, Sirkka L. Jarvenpaa, and Ann Majchrzak. “Knowledge collaboration in online communities”. en. In: *Organization science* 22.5 (2011), 1224–1239. ISSN: 1047-7039. DOI: 10.1287/orsc.1100.0614. URL: http://dx.doi.org/10.1287/orsc.1100.0614.

[31] Rosta Farzan and Robert E Kraut. “Socializing volunteering: designing to entice, entice, and retain”. In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 2010, pp. 167–176.

[32] Robert Feldt et al. “Validity threats in empirical software engineering research: An initial survey”. In: *Proceedings of the 10th International Conference on Software Process (ICSP)* (2010), pp. 15–28.

[33] Susan T Fiske. *Social beings: Core motives in social psychology*. John Wiley & Sons, 2018.

[34] Denae Ford, Alisse Harkins, and Chris Parnin. "Someone like me: How does peer parity influence participation of women on stack overflow?" In: *VL/HCC*. 2017, pp. 239–243. DOI: 10.1109/VLHCC.2017.8103473.

[35] Imane Fouad et al. "On Compliance of Cookie Purposes with the Purpose Specification Principle". In: *2020 IEEE European Symposium on Security and Privacy Workshops (EuroSPW)*. 2020, pp. 326–333. DOI: 10.1109/EuroSPW51379.2020.00051.

[36] Floyd J. Fowler. *Survey research methods*. en. 5th ed. Christchurch, New Zealand: Sage Publications, 2013. ISBN: 9781483323596.

[37] A. Michael Froomkin. "From anonymity to identification". In: *Journal of Self-Regulation and Regulation* 1 (2015). DOI: 10.11588/josar.2015.0.23480. URL: https://repository.law.miami.edu/cgi/viewcontent.cgi?article=1065&context=fac_articles.

[38] Adam Geitgey. *Face Recognition by Python*. en. 2023. URL: https://pypi.org/project/face-recognition.

[39] Inc GitHub. *GitHub REST API documentation*. en. 2022. URL: https://docs.github.com/en/rest.

[40] Torleif Halkjelsvik, Magne Jørgensen, and Karl Halvor Teigen. "To read two pages, I need 5 minutes, but give me 5 minutes and I will read four: how to change productivity estimates by inverting the question". en. In: *Applied cognitive psychology* 25.2 (2011), 314–323. ISSN: 0888-4080. DOI: 10.1002/acp.1693. URL: http://dx.doi.org/10.1002/acp.1693.

[41] Thomas Herrmann, Isa Jahnke, and Kai-Uwe Loser. "The role concept as a basis for designing community systems." In: *Coop*. 2004, pp. 163–178.

[42] Dwight M. Hite, Troy Voelker, and Adrian Robertson. "Measuring perceived anonymity: The development of a context independent instrument". In: *Journal of methods and measurement in the social sciences* 5.1 (2014), p. 22. ISSN: 2159-7855. DOI: 10.2458/jmm.v5i1.18305. URL: http://dx.doi.org/10.2458/jmm.v5i1.18305.

[43] Nguyen Phong Hoang and Davar Pishva. "Anonymous communication and its importance in social networking". In: *16th International Conference on Advanced Communication Technology*. 2014, pp. 34–39. DOI: 10.1109/ICACT.2014.6778917.

[44] Jaap-Henk Hoepman. "Privacy Design Strategies". In: *ICT Systems Security and Privacy Protection*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, 446–459. ISBN: 9783642554148.

[45] Timo Hynninen, Antti Knutas, and Maija Hujala. "What can we learn from recommendations of early-career engineers? Assessing computing and software engineering education using a career monitoring survey". In: *Proceedings of the 2022 Conference on United Kingdom & Ireland Computing Education Research*. UKICER '22. Dublin, Ireland: Association for Computing Machinery, 2022. ISBN: 9781450397421. DOI: 10.1145/3555009.3555017. URL: https://doi.org/10.1145/3555009.3555017.

[46] GitHub Inc. *GitHub General Privacy Statement*. URL: https://docs.github.com/en/site-policy/privacy-policies/github-general-privacy-statement (visited on 01/25/2024).

[47] Stack Exchange Inc. *Stack Overflow Privacy Statement*. URL: https://stackoverflow.com/legal/privacy-policy (visited on 01/24/2024).

[48] Corey Brian Jackson, Kevin Crowston, and Carsten Østerlund. "Did they login? Patterns of Anonymous Contributions in Online Communities". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW (2018). DOI: 10.1145/3274346. URL: https://doi.org/10.1145/3274346.

[49] Eric Jardine. "Tor, what is it good for? Political repression and the use of online anonymity-granting technologies". In: *New Media & Society* 20 (2018), pp. 435 – 452. URL: https://api.semanticscholar.org/CorpusID:4984943.

[50] Barbara Kitchenham and Shari Lawrence Pfleeger. "Principles of survey research part 4: Questionnaire evaluation". en. In: *Software engineering notes* 27.3 (2002), 20–23. ISSN: 0163-5948. DOI: 10.1145/638574.638580. URL: http://dx.doi.org/10.1145/638574.638580.

[51] Barbara Kitchenham and Shari Lawrence Pfleeger. "Principles of survey research: Part 5: populations and samples". en. In: *Software engineering notes* 27.5 (2002), 17–20. ISSN: 0163-5948. DOI: 10.1145/571681.571686. URL: http://dx.doi.org/10.1145/571681.571686.

[52] Barbara Kitchenham and Shari Lawrence Pfleeger. "Principles of survey research part 6: Data analysis". en. In: *Software engineering notes* 28.2 (2003), 24–27. ISSN: 0163-5948. DOI: 10.1145/638750.638758. URL: http://dx.doi.org/10.1145/638750.638758.

[53] Barbara Kitchenham et al. "Robust statistical methods for empirical software engineering". In: *Empirical Software Engineering* 22 (2017), pp. 579–630.

[54] Barbara A Kitchenham, Tore Dyba, and Magne Jorgensen. "Evidence-based software engineering and systematic reviews". In: *Proceedings of the 30th International Conference on Software Engineering (ICSE)*. ACM. 2008, pp. 273–281.

[55] Barbara A. Kitchenham and Shari Lawrence Pfleeger. "Principles of survey research part 2: Designing a survey". en. In: *Software engineering notes* 27.1 (2002), 18–20. ISSN: 0163-5948. DOI: 10.1145/566493.566495. URL: http://dx.doi.org/10.1145/566493.566495.

[56] Barbara A. Kitchenham and Shari Lawrence Pfleeger. "Principles of survey research: Part 3: constructing a survey instrument". en. In: *Software engineering notes* 27.2 (2002), 20–24. ISSN: 0163-5948. DOI: 10.1145/511152.511155. URL: http://dx.doi.org/10.1145/511152.511155.

[57] Umme Ayman Koana et al. "Examining Ownership Models in Software Teams: A Systematic Literature Review and a Replication Study". In: *arXiv preprint arXiv:2405.15665* (2024).

[58] Umme Ayman Koana et al. "Ownership in the hands of accountability at brightsquid: A case study and a developer survey". In: *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 2023, pp. 2008–2019.

[59] Hong Joo Lee et al. "Anonymous participation and collaboration efficiency in online communities". In: *Asia Pacific journal of information systems* 30.3 (2020), 497–512.

ISSN: 2288-5404. DOI: 10.14329/apjis.2020.30.3.497. URL: http://dx.doi.org/10.14329/apjis.2020.30.3.497.

[60] Roy Ka-Wei Lee and David Lo. "GitHub and stack overflow: Analyzing developer interests across multiple social collaborative platforms". In: *Lecture Notes in Computer Science*. Cham: Springer International Publishing, 2017, 245–256. ISBN: 9783319672557.

[61] Bin Lin and Alexander Serebrenik. "Recognizing gender of stack overflow users". In: *MSR*. New York, NY, USA: ACM, 2016. ISBN: 9781450341868.

[62] Mario Linares-Vásquez et al. "Continuous integration in a social-coding world: Empirical evidence from GitHub". In: *Proceedings of the 2014 IEEE International Conference on Software Maintenance and Evolution*. IEEE. 2014, pp. 401–405.

[63] Yao Lu et al. "Motivation under gamification: An empirical study of developers' motivations and contributions in stack overflow". In: *TSE* (2021). ISSN: 0098-5589. DOI: 10.1109/tse.2021.3130088. URL: http://dx.doi.org/10.1109/tse.2021.3130088.

[64] Mac Malone et al. "To Gamify or Not? On Leaderboard Effects, Student Engagement and Learning Outcomes in a Cybersecurity Intervention". In: *ACM Technical Symposium on Computer Science Education*. SIGCSE '21. Virtual Event, USA: Association for Computing Machinery, 2021, 1135–1141. ISBN: 9781450380621. DOI: 10.1145/3408877.3432544. URL: https://doi.org/10.1145/3408877.3432544.

[65] Jennifer Marlow and Laura Dabbish. "Activity traces and signals in software developer recruitment and hiring". In: *conference on Computer supported cooperative work*. 2013, pp. 145–156.

[66] Gary T. Marx. "What's in a name? Some reflections on the sociology of anonymity". In: *The information society* 15.2 (1999), 99–112. ISSN: 0197-2243. DOI: 10.1080/019722499128565. URL: http://dx.doi.org/10.1080/019722499128565.

[67] Célestin Matte, Nataliia Bielova, and Cristiana Santos. "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)*. 2020, pp. 791–809. DOI: 10.1109/SP40000.2020.00076.

[68] Poppy Lauretta McLeod. "Effects of Anonymity and Social Comparison of Rewards on Computer-Mediated Group Brainstorming". In: *Small Group Research* 42 (2011), pp. 475 –503. URL: https://api.semanticscholar.org/CorpusID:145140476.

[69] Jill Mosteller and Amit Poddar. "To share and protect: Using regulatory focus theory to examine the privacy paradox of consumers' social media engagement and online privacy protection behaviors". en. In: *Journal of interactive marketing* 39 (2017), 27–38. ISSN: 1094-9968. DOI: 10.1016/j.intmar.2017.02.003. URL: http://dx.doi.org/10.1016/j.intmar.2017.02.003.

[70] Emerson Murphy-Hill, Thomas Zimmermann, and Nachiappan Nagappan. "Cowboys, ankle sprains, and keepers of quality: how is video game development different from software development?" In: *Proceedings of the 36th International Conference on Software Engineering*. New York, NY, USA: ACM, 2014. ISBN: 9781450327565.

[71] Kumiyo Nakakoji et al. "Evolution patterns of open-source software systems and communities". In: *Proceedings of the international workshop on Principles of software evolution*. ACM. 2002, pp. 76–85.

[72] Maleknaz Nayebi. "Eye of the mind: Image processing for social coding". In: *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results*. 2020, pp. 49–52.

[73] Maleknaz Nayebi, Bram Adams, and Guenther Ruhe. "Release Practices for Mobile Apps–What do Users and Developers Think?" In: *2016 ieee 23rd international conference on software analysis, evolution, and reengineering (saner)*. Vol. 1. IEEE. 2016, pp. 552–562.

[74] Maleknaz Nayebi, Guenther Ruhe, and Thomas Zimmermann. "Mining treatment-outcome constructs from sequential software engineering data". In: *IEEE Transactions on Software Engineering* 47.2 (2019), pp. 393–411.

[75] Maleknaz Nayebi et al. "A longitudinal study of identifying and paying down architectural debt". 2018. URL: http://arxiv.org/abs/1811.12904.

[76] Maleknaz Nayebi et al. "Analytics for Software Project Management–Where are We and Where do We Go?" In: *2015 30th IEEE/ACM International Conference on Automated Software Engineering Workshop (ASEW)*. IEEE. 2015, pp. 18–21.

[77] Maleknaz Nayebi et al. "Anatomy of functionality deletion: an exploratory study on mobile apps". In: *Proceedings of the 15th International Conference on Mining Software Repositories*. 2018, pp. 243–253.

[78] Maleknaz Nayebi et al. "ESSMArT way to manage customer requests". In: *Empirical Software Engineering* 24 (2019), pp. 3755–3789.

[79] Maleknaz Nayebi et al. "Hybrid labels are the new measure!" In: *IEEE Software* 35.1 (2017), pp. 54–57.

[80] Maleknaz Nayebi et al. "Recommending and release planning of user-driven functionality deletion for mobile apps". In: *Requirements Engineering* (2024), pp. 1–22.

[81] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. en. Palo Alto, CA: Stanford University Press, 2010. ISBN: 9780804752374.

[82] Helen Nissenbaum. "The meaning of anonymity in an information age". In: *The information society* 15.2 (1999), 141–144. ISSN: 0197-2243. DOI: 10.1080/019722499128592. URL: http://dx.doi.org/10.1080/019722499128592.

[83] Payod Panda et al. "AllTogether: Effect of avatars in mixed-modality conferencing environments". In: *Symposium on Human-Computer Interaction for Work*. New York, NY, USA: ACM, 2022.

[84] Hyunjung Park and Sung Joo Park. "Communication behavior and online knowledge collaboration: evidence from Wikipedia". en. In: *Journal of knowledge management* 20.4 (2016), 769–792. ISSN: 1367-3270. DOI: 10.1108/jkm-08-2015-0312. URL: http://dx.doi.org/10.1108/jkm-08-2015-0312.

[85] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. *Deep face recognition*. 2015.

[86] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. *Deep Face Recognition*. 2015. URL: https://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/parkhi15.pdf.

[87] Sai Teja Peddinti, Keith W. Ross, and Justin Cappos. “"On the internet, nobody knows you’re a dog": a twitter case study of anonymity in social networks”. In: *ACM Conference on Online Social Networks*. COSN ’14. Dublin, Ireland: Association for Computing Machinery, 2014, 83–94. ISBN: 9781450331982. DOI: 10.1145/2660460.2660467. URL: https://doi.org/10.1145/2660460.2660467.

[88] Shari Lawrence Pfleeger and Barbara A. Kitchenham. “Principles of survey research: Part 1: turning lemons into lemonade”. en. In: *Software engineering notes* 26.6 (2001), 16–18. ISSN: 0163-5948. DOI: 10.1145/505532.505535. URL: http://dx.doi.org/10.1145/505532.505535.

[89] Stephen A. Rains. “The impact of anonymity on perceptions of source credibility and influence in computer-mediated group communication: A test of two competing hypotheses”. en. In: *Communication research* 34.1 (2007), 100–125. ISSN: 0093-6502. DOI: 10.1177/0093650206296084. URL: http://dx.doi.org/10.1177/0093650206296084.

[90] Gema Rodríguez-Pérez, Reza Nadri, and Meiyappan Nagappan. “Perceived diversity in software engineering: a systematic literature review”. en. In: *Empirical Software Engineer* 26.5 (2021), p. 102. ISSN: 1382-3256. DOI: 10.1007/s10664-021-09992-2. URL: http://dx.doi.org/10.1007/s10664-021-09992-2.

[91] Günther Ruhe and Maleknaz Nayebi. “What counts is decisions, not numbers—toward an analytics design sheet”. In: *Perspectives on Data Science for Software Engineering*. Elsevier, 2016, pp. 111–114.

[92] Thais Sardá et al. “Understanding online anonymity”. en. In: *Media, culture, and society* 41.4 (2019), 557–564. ISSN: 0163-4437. DOI: 10.1177/0163443719842074. URL: http://dx.doi.org/10.1177/0163443719842074.

[93] Anita Sarma et al. “Hiring in the Global Stage: Profiles of Online Contributions”. In: *ICGSE*. 2016, pp. 1–10. DOI: 10.1109/ICGSE.2016.35.

[94] Sk Golam Saroar and Maleknaz Nayebi. “Developers’ perception of GitHub Actions: A survey analysis”. In: *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*. 2023, pp. 121–130.

[95] L. Sayrs. "InterViews: An introduction to qualitative research interviewing Steinar Kvale. Thousand Oaks, CA: Sage, 1996. 326 pp". en. In: *The American journal of evaluation* 19.2 (1998), 267–270. ISSN: 1098-2140. DOI: 10.1016/s1098-2140(99)80208-2. URL: http://dx.doi.org/10.1016/s1098-2140(99)80208-2.

[96] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A unified embedding for face recognition and clustering". In: *CVPR*. 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.

[97] Sefik Ilkin Serengil and Alper Ozpinar. "LightFace: A Hybrid Deep Face Recognition Framework". In: *ASYU*. 2020, pp. 1–5. DOI: 10.1109/ASYU50717.2020.9259802.

[98] Donghee Shin. "User Perceptions of Algorithmic Decisions in the Personalized AI System:Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability". In: *Journal of Broadcasting & Electronic Media* 64 (2020), pp. 541 – 565. URL: https://api.semanticscholar.org/CorpusID:230586657.

[99] Yan Shvartzshnaider et al. "Going against the (appropriate) flow: A contextual integrity approach to privacy policy analysis". In: *AAAI Conference on Human Computation and Crowdsourcing* 7 (2019), 162–170. ISSN: 2769-1349. DOI: 10.1609/hcomp.v7i1.5266. URL: http://dx.doi.org/10.1609/hcomp.v7i1.5266.

[100] Vivek Kumar Singh et al. "Anonymous developers: Who are they and do they matter?" In: *Journal of Systems and Software* 156 (2019), pp. 352–364.

[101] Sphinx. *Structural Similairty Index*. en. 2013. URL: https://scikit-image.org/docs/stable/auto_examples/transform/plot_ssim.html.

[102] Peter Suciu. *Trolls continue to be A problem on social media*. en. 2020. URL: https://www.forbes.com/sites/petersuciu/2020/06/04/trolls-continue-to-be-a-problem-on-social-media/?sh=3e5d34323a89.

[103] Yi Sun, Xiaogang Wang, and Xiaoou Tang. "Deep learning face representation from predicting 10,000 classes". In: *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014. ISBN: 9781479951185.

[104] Yaniv Taigman et al. "DeepFace: Closing the gap to human-level performance in face verification". In: *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014. ISBN: 9781479951185.

[105] Martin Tanis and Tom Postmes. "Two faces of anonymity: Paradoxical effects of cues to identity in CMC". en. In: *Computers in human behavior* 23.2 (2007), 955–970. ISSN: 0747-5632. DOI: 10.1016/j.chb.2005.08.004. URL: http://dx.doi.org/10.1016/j.chb.2005.08.004.

[106] Welderufael B. Tesfay et al. "I Read but Don't Agree: Privacy Policy Benchmarking using Machine Learning and the EU GDPR". In: *Companion Proceedings of the The Web Conference 2018*. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, 2018, 163–166. ISBN: 9781450356404. DOI: 10.1145/3184558.3186969. URL: https://doi.org/10.1145/3184558.3186969.

[107] Christoph Treude. "Qualitative data analysis in software engineering: Techniques and teaching insights". 2024. URL: http://arxiv.org/abs/2406.08228.

[108] UOM. *Labeled Faces in the Wild Dataset*. en. 2013. URL: http://vis-www.cs.umass.edu/lfw/.

[109] Asimina Vasalou and Adam N Joinson. "Me, myself and I: The role of interactional context on self-presentation through avatars". In: *Computers in human behavior* 25.2 (2009), pp. 510–520.

[110] Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. "StackOverflow and GitHub: Associations between software development and crowdsourced knowledge". In: *2013 International Conference on Social Computing*. IEEE, 2013. ISBN: 9780769551371.

[111] Bogdan Vasilescu et al. "Gender and tenure diversity in GitHub teams". In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM. 2014, pp. 3789–3798.

[112] Bogdan Vasilescu et al. "How social QA sites are changing knowledge sharing in open source software communities". In: *ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '14. Baltimore, Maryland, USA: Association for Computing Machinery, 2014, 342–354. ISBN: 9781450325400. DOI: 10.1145/2531602.2531659. URL: https://doi.org/10.1145/2531602.2531659.

[113] Kathleen D Vohs, Roy F Baumeister, and Natalie J Ciarocco. "Self-regulation and self-presentation: regulatory resource depletion impairs impression management and effortful self-presentation depletes regulatory resources." In: *Journal of personality and social psychology* 88.4 (2005), p. 632.

[114] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

[115] Xiahua Wei, Wei Chen, and Kevin Zhu. "Motivating user contributions in online knowledge communities: Virtual rewards and reputation". In: *Conference on System Sciences*. IEEE, 2015. ISBN: 9781479973675.

[116] Igor Scaliante Wiese et al. "Who is who in the mailing list? Comparing six disambiguation heuristics to identify multiple addresses of a participant". In: *ICSME*. IEEE, 2016. ISBN: 9781509038060.

[117] Hana Winchester, Alicia E. Boyd, and Brittany Johnson. "An exploration of intersectionality in software development and use". In: *Proceedings of the Third Workshop on Gender Equality, Diversity, and Inclusion in Software Engineering*. New York, NY, USA: ACM, 2022.

[118] Claes Wohlin et al. "Experimentation in software engineering". In: *Springer Science & Business Media* (2012).

[119] Claes Wohlin et al. *Experimentation in Software Engineering: An Introduction*. en. New York, NY: Springer, 2012. ISBN: 9781461370918. DOI: 10.1007/978-1-4615-4625-2. URL: http://dx.doi.org/10.1007/978-1-4615-4625-2.

[120] Lior Wolf, Tal Hassner, and Itay Maoz. "Face recognition in unconstrained videos with matched background similarity". In: *CVPR 2011*. IEEE, 2011. ISBN: 9781457703942.

[121] Richmond Y. Wong and Deirdre K. Mulligan. "Bringing Design to the Privacy Table: Broadening "Design" in "Privacy by Design" Through the Lens of HCI". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, 1–17. ISBN: 9781450359702. DOI: 10.1145/3290605.3300492. URL: https://doi.org/10.1145/3290605.3300492.

[122] Corp. X. *Twitter User account metadata API*. en. 2013. URL: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/user.

[123] Yunxiang Xiong et al. "Mining developer behavior across GitHub and StackOverflow". In: *Conference on Software Engineering and Knowledge Engineering*. KSI Research Inc. and Knowledge Systems Institute Graduate School, 2017. ISBN: 9781891706417.

[124] Lei Xu, Tingting Nian, and Luís Cabral. "What makes geeks tick? A study of Stack Overflow careers". en. In: *Management science* 66.2 (2020), 587–604. ISSN: 0025-1909. DOI: 10.1287/mnsc.2018.3264. URL: http://dx.doi.org/10.1287/mnsc.2018.3264.

[125] İlke Yurtseven, Selami Bagriyanik, and Serkan Ayvaz. "A review of spam detection in social media". In: *Conference on Computer Science and Engineering (UBMK)*. IEEE. 2021, pp. 383–388.

[126] Dujuan Zhang, Jie Li, and Zhenfang Shan. *Implementation of dlib deep learning face recognition technology*. IEEE, 2020.

[127] Kaipeng Zhang et al. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks". In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503. DOI: 10.1109/LSP.2016.2603342.

[128] Xin Zhao and Narissa Tsuboi. "Early career software developers - are you sinking or swimming?" In: *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Society*. New York, NY, USA: ACM, 2024.

[129] Wenhan Zhu et al. "An empirical study of question discussions on Stack Overflow". en. In: *Empirical Software Engineer* 27.6 (2022). ISSN: 1382-3256. DOI: 10.1007/s10664-022-10180-z. URL: https://plg.uwaterloo.ca/~migod/papers/2022/emse22-chatRooms-officlalPreprint.pdf.

[130] Sebastian Zimmeck and Steven M. Bellovin. "Privee: An Architecture for Automatically Analyzing Web Privacy Policies". In: *23rd USENIX Security Symposium (USENIX Security 14)*. San Diego, CA: USENIX Association, Aug. 2014, pp. 1–16. ISBN: 978-1-931971-15-7. URL: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/zimmeck.

Appendix A: Interview Study Questions

Goal of this interview survey:

- Understanding the degree of information sharing and the sensitivity of information sharing for users on social coding platforms with regard to different identity factors.
- Determining the intent of the users of social coding platforms for anonymity.
- Determining the balance cost model of anonymity in comparison to professional benefits and visibility.

Demographics

The behavior of developers on social coding platforms is impacted by variety of demographic, cultural, and personal factors. In this part, we would like to ask for your demographic information helping us to contextualize the research results. These information (and any other information in this interview) would not be used individually and will only be reported in aggregation. We will treat this information with utmost confidentiality and discard these responses after aggregation and reporting. Participation in this interview is voluntary and you may leave the interview at any point desired.

1. What is your current professional role or job title?
2. What province (state) do you currently reside?
   - Alberta
   - British Columbia
   - Manitoba
   - New Brunswick
   - Newfoundland and Labrador

- Nova Scotia
- Ontario
- Prince Edward Island (PEI)
- Quebec
- Saskatchewan
- Northwest Territories, Nunavut, or Yukon
- Other (please specify here):
- Prefer not to answer.

3. How many years of experience do you have in software development?

- 0-1 year
- 1-2 years
- 2-3 years
- 3-4 years
- 4-5 years
- More than 5 years
- Prefer not to answer

4. What is your age range?

- Under 18
- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65 and over
- Prefer not to answer

5. What is your self-identified gender?

- Male
- Female
- Non-binary
- Other (please specify)
- Prefer not to answer

6. What is your sexual orientation?

- Asexual
- Bisexual
- Gay
- Lesbian
- Queer
- Questioning
- Straight (Heterosexual)
- Two-Spirit
- Pansexual
- Other (please specify)
- Prefer not to answer

7. What is your Race/Ethnicity?

- White (e.g., German, Irish, English, Italian, Polish, French, etc.)
- Hispanic, Latino, or Spanish origin (e.g., Mexican or Mexican American, Puerto Rican, Cuban, Salvadoran, Dominican, Colombian, etc.)
- Black or African American (e.g., African American, Jamaican, Haitian, Nigerian, Ethiopian, Somalian, etc.)
- Asian (e.g., Chinese, Filipino, Asian Indian, Vietnamese, Korean, Japanese, etc.)
- American Indian or Alaska Native (e.g., Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Native Village or Barrow Inupiat Traditional Government, Nome Eskimo Community, etc.)

- Middle Eastern or North African (e.g., Lebanese, Iranian, Egyptian, Syrian, Moroccan, Algerian, etc.)
- Native Hawaiian or Other Pacific Islander (e.g., Native Hawaiian, Samoan, Chamorro, Tongan, Fijian, etc.)
- Other race, ethnicity, or origin (please specify):
- Prefer not to answer

8. Do you identify as Indigenous?

- Yes. (please specify)
  - First Nations
  - Metis
  - Inuit
  - Other (please specify)
- No
- Prefer not to answer

9. Do you consider yourself as having a disability?

- Yes
- No
- Prefer not to answer

10. Can you provide us with your Stack Overflow and GitHub username?

Intent-Based Questions (of the Individual)

1. What social media apps do you use?

2. Can you please list the social coding platforms and professional networks that you use, specifically relevant to your profession?

3. How often (on average) do you update your information on social coding profiles (On a scale of 1 (being rarely) to 5 (being very frequently and as any change is happening)?

4. Do you use your real photo on your social media? What about on the social coding platforms and professional networks you mentioned?

5. What social media, social coding profiles, professional networks have you included in your resume or CV and why?

6. Do you have multiple accounts on your social media apps or social coding platforms? Can you explain your reasons behind whether or not maintaining multiple accounts on these platforms?

7. How would you describe being anonymous online?

8. Do you think that having multiple accounts assists with your anonymity on these platforms?

9. Are you anonymous based on the definition we discussed on any of your social media, social coding platforms or professional network accounts? (If yes) Can you elaborate on the specific reasons for maintaining anonymity on these platforms?

10. What profile information are the most important for you to protect your anonymity? why?

11. Can you elaborate on situations or interactions (activities) where you find anonymity to be crucial on these platforms?

12. How important is it for you to remain anonymous on these platforms (on a scale of 1, not important, to 5, extremely important)?

13. Do you have verified accounts on the platforms in any of your social media, social coding platforms or professional networks? if yes which ones, and if not why?

14. If it was required to verify your Stack Overflow and GitHub profiles how would that impact your activity on each platform?

15. Can you discuss how the anonymity of others influences your decision in terms of sharing information on platforms or engaging with them?

Privacy Policy Based Questions

1. Which of the privacy policies, either Stack Overflow's or GitHub's, have you read?

2. (if 1 yes) To what extent have you read the privacy polices in depth to understand the implications for your personal information and content? While 1 being skimming the policy and 5 being in all details.

3. (if 1 yes) Currently, to what extent do you remember each policy (for Stack Overflow or GitHub)?

   - Not at all

- Vaguely
- Moderately
- Quite well
- Extremely well

4. (if 1 is no jump here) Can you share the reasons on why you have not read these privacy policies?

5. For Stack Overflow we could retrieve 12 pages of privacy policy on personal data. Here I have a print. How much time would it take you to read this? ….. as we are doing a controlled study I am going to ask you read this document to the best you can in 10 minutes and highlight any parts that are most important for you. After the 10 min is over…., Given now that you have read it:

6. How likely is it for you to indeed read through this policy fully (1 being not likely at all and 5 being extremely likely)?

7. For you, what was the most important policy in Stack Overflow's privacy policy?

8. For GitHub we could retrieve 5 pages of privacy policy on personal data. Here I have a print. How much time would it take you to read this? ….. as we are doing a controlled study I am going to ask you read this document to the best you can in 5 minutes and highlight any parts that are most important for you. After the 5 min is over…., Given now that you have read it:

9. How likely is it for you to indeed read through this policy fully? Why this more likely than when you read stack overflow's privacy policy?

10. For you, was the most important policy in GitHub's Privacy Policy? (1 being not likely at all and 5 being extremely likely)?

To make sure that the users understand the privacy policy, we are investigating the use of contextual integrity (CI). We restructured these policies using CI. The theory of privacy as contextual integrity (CI) introduces three key concepts into the privacy vocabulary: information flow (sender, subject, recipient, attribute, transmission principle(reason).

11. Here is the table of the privacy policy using CI for Stack Overflow, after manually annotating them using Contextual Integrity from the section of the privacy policy which was coded under Personal data. How would you rank these in order of importance in terms of which ones would you care to be notified of first, second and so on?

12. Here is the table of the privacy policy using CI for GitHub after manually annotating them using Contextual Integrity from the section of the privacy policy which was coded under Personal data. How would you rank these in order of importance in terms of which ones would you care to be notified of first, second and so on?

13. Do you believe adding the CI structure helped you with understanding of these policies?

14. How likely would you read these policies if they were structured as such(1 being not likely at all and 5 being extremely likely)?

- Before we wrap up we wanted to ask if you would be interested in us following up with you with a report of how anonymous or how findable you are on the social coding platforms and professional networks.
- Can you please fill in the demographics?
- Thank you for your time!

Appendix B: Stack Overflow Privacy Policy

This appendix includes Stack Overflow's privacy policy statement [47].

The Stack Overflow Network (also referred to herein as "Stack Overflow" or "Network") is a set of related Internet sites and other applications for asynchronous collaboration and sharing knowledge, owned and operated by Stack Exchange, Inc. ("Stack Overflow", "we" or "us"), a Delaware corporation. Stack Overflow welcomes you to the Network, the largest community of developers and technologists in the world, and invites you to participate in the community by sharing knowledge with your peers and colleagues. WE DO NOT AND WILL NOT SELL YOUR PERSONAL INFORMATION.

INTRODUCTION TO OUR PRIVACY POLICY

Stack Overflow's products, services and websites can be accessed by users across various locations including in jurisdictions which have specific privacy requirements. This privacy policy and our related Privacy Notices (see below) describe how we collect, use, protect, share and transfer information about you that we collect when you use any of our products and services, our websites, or apps or otherwise interact with us or with our Network (Services). It also includes information about how we handle information on California residents under the CCPA. "Personal data" in this policy, includes "personal information," as it is defined under the CCPA. Personal data is any information about you by which you can be identified or be identifiable (either on its own or when combined with other information). This can include information such as: your name, email address and username; information about your device (e.g., IP address); and information relating to how you use and interact with our sites, apps and services. We have set out below an overview of how our privacy notices work. Depending on how you interact with us, we will collect and process your personal information as shown below. This Privacy Notice applies where you interact with us through the Network, or

use any of our products and services. If you interact with us through our Collectives on Stack Overflow product, please read the Collectives Privacy Notice. If you interact with us through your organization's Private Network on Teams Free, Basic, Business or Enterprise, we will only process certain limited information as controller. For Teams Free, Basic, Business: Please read the . Privacy Notice for Stack Overflow Teams Free, Basic and Business For Teams Enterprise: Please read the . Privacy Notice for Stack Overflow Teams Enterprise If you're an employee, our Employee Privacy Notice will be applicable. Please read the Privacy Notice for Stack Overflow Talent and Jobs.

GENERAL PRIVACY NOTICE FOR STACK EXCHANGE, INC.

This Privacy Notice is applicable to you, including when you participate in our Public Community Platform for Questions and Answers by interacting with us through Stack Overflow's Public Network ("Public Network"), or when you otherwise interact with us, e.g., when you purchase or use our Advertising or Teams or Collectives or other products ("Products") or when you contact our Customer Service, Sales or other teams for assistance or inquiries ("Services"). The Public Network allows individual users to engage in a community to share their knowledge and expertise, and to acquire knowledge and expertise. It is communitymoderated, and the content in the Public Network is public and therefore visible for anyone to see and is subject to our Public Network Terms of Use. Below you will find details about the personal information we collect about you when you use the Public Network or our Products and Services and interact with us; how and why we collect it, how we use it, and how we keep it secure. This Privacy Notice also tells you about your rights, how you can exercise those rights, and how the law protects you. Stack Exchange, Inc. is the Controller responsible for your personal data collected on the Public Network or directly by us for Stack's operational purposes. In connection with certain Products and Services, Stack is the Processor for the organizations who provide your data to us or use our Products and Services. If you interact with Stack Overflow through your organization, please refer to the specific privacy notices applicable to those Products. If you have any questions about this Privacy Notice or the way in which we use your personal data, please contact us at privacy@stackoverflow.com WHAT INFORMATION DO WE COLLECT? This section sets out the categories and details of personal information that we collect across our Products and Services. We collect and process your personal information in order to offer the Network, and to offer you our Products and

Services. If you decline to provide us with personal information when we ask you, you may not be able to access all or some of our services, and you may not be able to use the Network or certain Products and Services. For example, if you do not provide us with your email address or other information required for account registration, we will not be able to register you for certain account services. We collect: Identity Data We collect information such as your username, password, email address, IP address (collected at each site visit). This information is also used to identify individual users to the Public Network and to award activities and attribute them to you. Your username and additional content you provide, including your profile picture, question and answer content, and your reputation, are visible publicly. We or our community moderators may delete a question, an answer, a comment, or all of them and after deletion such information is only available to certain community members or Stack employees.

Contact Data: including delivery address, email address, and telephone numbers for operational purposes, including providing Services and Products. Technical Data: including internet protocol (IP) address, your login data, traffic data, browser type and version, time zone setting and location, browser plug-in types and versions, operating system, and platform and other communication data which gives us information about how you accessed our website. Other account information and usage information including your IP address and browser data may be used for diagnosis, security, and Product and Service optimization. Account registration information When you register for a Stack Overflow account, we collect your name and email address, your username and any additional information you provide. We will generate a unique identification number that we will use to recognize you when you are signed in. We need to collect your personal information for account registration purposes. Although the Network provides for anonymous and pseudonymous participation, in some instances in order to use certain Products or Services, we require account registration. This requires a name associated with your account, an email address at which we can contact you, and in some cases, additional information, including a contact address, a billing address, and a password to help secure your personal information. Billing information For Products and Services requiring payment, we collect a billing name, phone number, and email address. We also collect a billing and shipping address for invoicing purposes. If you elect to pay by credit card, Stack Overflow may engage a third party to securely process your payment.

All payment processing is done through a PCI DSS compliant third party. If customers are paying by a credit card, their card details are vaulted through a third-party payment gateway. Stack Overflow will store an encrypted token along with the last four digits of the credit card and the expiration month and year of the card and will not store or retain any other billing information about you.

Profile information

When completing a public "profile", whether it is for use with the Network or our Products and Services, we may require you to share certain personally identifying information in required and optional profile fields. The name associated with your account, which you may review and change in your account settings, as well as reputation points are publicly displayed and connected to your activity on the Stack Overflow Network and via certain Products and Services, such as our Talent and Teams platforms. Moderators, who have elevated access privileges, may access nonpublic user profile data, as well as content in posts and other public data. For more information about moderators and their role in the Stack Exchange Network, please visit our help center. Other profile information may also be publicly displayed if you choose to complete certain elective components of the Account Registration options, including e.g., your Developer Story, display picture, affiliated role and company, your location, and your preferred technologies (javascript, swift, sql-server, etc.), and some information is publicly displayed and required, such as how long you have been a member. Similarly, you may elect to share additional information that is not required, but this will be displayed publicly if you so choose to share it. If you are using the private Network (e.g., Stack Overflow for Teams), your account settings, including name, affiliated role and company, and other information may be shared with your private Network Team members designated as authorized users of your private network account. While some of the public and private Network profile information is required, much of it is within your control as we describe herein and can be modified, updated, and restricted by you in your settings. Information we use to identify and authenticate you We need this information to be able to recognize you, for example, when you access your accounts. It also helps us to protect your details and keep them secure. We may also use this information for fraud detection and prevention. Analytics information Stack Overflow uses data analytics to ensure site functionality and to optimize our Product and Service offerings to you. We use web browser and mobile

analytics to allow us to understand Network and Apps functionality. In doing so, we record information including, for example, how often you visit the Network, how often you contribute content, Network and Apps performance data, errors and debugging information, and the type of activity you engage in while on the Network or in your use of our Products and Services. We may on occasion share this information with third parties with whom we have contracts for products and services to assist Stack. Information from Developer surveys, questionnaires, research and feedback programs We collect information through questionnaires, surveys and feedback programs to help improve our products and give us insights. We may also conduct similar research for advertisers and our marketing partners. We ask you for your consent to use this information when you participate in these programs and events. Marketing and sales information We collect details of the services you receive and your preferences; information about your device or the software you use, e.g., its IP address, technical specification and uniquely identifying data; cookies and similar technologies we use to recognize you, remember your preferences and tailor the content we provide to you – our contains more details about how we use cookies. cookie policy Device and browser information received automatically When you visit the Network or use our Apps, Stack Overflow automatically receives and records information from your browser or mobile device, such as your Internet Protocol (IP) address or unique device identifier. Cookies and data about which pages you visit on our Network allow us to operate and optimize the Products and Services we provide to you. This information is stored in secure logs and is collected automatically. We may combine this browser information with other information we collect about you. This information is used to keep the Products and Services secure, to analyze and understand how our Products and Services are used, optimize such usage, provide advertising across the Network as well as certain Products and Services to personalize your experience, and to help connect you with potential job opportunities in the case of our recruiting Products and Services. Location information When you use the Stack Overflow Network, and certain of our Products and Services, we collect location information about you, including your IP address, your location, browser information, and how you came to the Stack Overflow Network. This is the case for individuals who have registered for an account, and non-members who engage with the Stack Overflow Network by visiting our website(s), but who have not completed an account registration. This information is collected

for various purposes, including advertising, analytics and to serve content as it relates to certain Products or Services (e.g., Talent), as well as to provide you with localized content, recommendations, and marketing. You may revoke our permission to collect some of this data, including your location and browser information through your Account Settings, but this may limit functionality in some cases. Certain location information we collect is required for security and site functionality. We share this information with certain third-parties (e.g., talent recruiters, payment processors, and advertising providers) in order to provide you with our Products and Services. You should not include any financial information or other information that you do not wish to make public when using our Public Network, which is a public website. We do not collect such information. It is your responsibility to keep such information safe and secure. Can I revoke permissions once I have put information on the Public Network? Please be aware that when using the Public Question and Answer Network, you cannot revoke permission for Stack Overflow to publish and to allow others to have derivative rights to such public content once it is made public.

The content you contribute on the public network is public, in perpetuity, and Stack Overflow loses direct control over such content when it is published on the open internet. This also means that if you post personal information about yourself, you do so at your own risk, including the risk of abandoning any privacy rights you had in such information. However, we can work with users to remove any personal data that has been posted in error or which a user wishes to remove. Additionally, from time to time, Stack Overflow makes such public content available via our Stack Overflow Network API, compilation via the Creative Commons Data Dump and by providing public content to the Stack Overflow Network. You should be aware that such content, once public, cannot often be removed from public view even if it is removed from the Stack Overflow network and our servers. How we collect information from you Information that you give to us

HOW WE USE YOUR PERSONAL DATA

This section explains how we use your information and what we need it for (purposes). We use your personal data only when we have a valid reason and the legal grounds to do so. We determine the legal grounds based on the purposes for which we have collected your personal data. More information is provided below: Purposes for

collecting and using personal information We use the above categories of personal information for the following purposes: To verify your identity and enable you to access your products and services when you contact us or access our services. To protect you from fraud and for fraud prevention and detection. To provide commercial quotes to you, e.g., when you inquire about a product or service. For statistical analysis (e.g., on the use of our websites). To operate and improve our websites and services. To notify you of any changes to our websites or our services and products that may affect you. To provide you with technical and customer support.

To ask your opinion or feedback on our services or surveys. To communicate with you about other Stack Overflow products or services, events, offers, promotions and other information or news that we think may be of interest to you. Whenever we communicate with you in this way, we will give you an opportunity to opt out of receiving future messages. To enforce our legal rights or comply with legal requirements. To provide improved website and product experience and communications informed by your product subscriptions and/or data collected. Our legal grounds for using your personal data We will only use your personal data when the law allows us to. We will use your personal data in the following circumstances: Where we need to perform the contract we are about to enter into or have entered into with you. For example, where you have purchased a product or service or requested a service from us. Where we have your consent. For example, where you have provided your consent to receive marketing emails from us or where you have given your consent to be contacted by a third party sponsor for an event you have participated in. You can withdraw your consent at any time. In the case of marketing emails, you can withdraw your consent by clicking on the "unsubscribe" link in the email you received. Where we need to comply with a legal or regulatory obligation. Where it is in our legitimate interests, including our commercial interests or a third party's legitimate interest in using the personal information. Examples include when we analyze what content has been viewed on our Network and apps, so that we can understand how they are used and improve our content; carrying out marketing analyses to better understand your interests and preferences so that we can make our marketing more relevant to your interests and preferences. This includes when we promote our own products and services. Whenever we rely on legitimate interests for processing, we ensure that we consider and balance our interests against the individual's interests

before processing. We do not use your personal data on the basis of legitimate interest for activities where individuals' interests override our interests. The table below gives further details of the personal data we collect and our reasons (legal basis) for using the information.

More details about how we use your personal data Cookies and Similar Technologies When you visit our Network and sites or when you use our apps, we may collect personal data from you automatically using cookies or similar technologies. A cookie is a small file that can be placed on your device that allows us to recognize and remember you. We may place cookies on your device (with your consent, where applicable), as well as partner with third parties, including advertising partners, who may use cookies or other similar technologies to provide you with advertising based on your browsing activities and interests. Please see our for additional information on how we use cookies and what third party cookies we permit. cookie policy Tracking Stack Overflow emails use Javascript for email tracking of opens, clicks, and unsubscribes or may contain a small, invisible, one pixel image which can be used to let us know when the email is opened or forwarded. If the email recipient has images enabled on their device, when the email is opened, a message (an automatic request) is sent to Stack Overflow's server requesting the image. That request is logged by Stack Overflow and tells us that the email has been opened. A similar process may be used in relation to links or attachments in the email, with a request being sent to Stack Overflow for the content that is accessible via the links or attachments, thus telling Stack Overflow that the link or attachment has been accessed. Developer Survey As the world's largest developer community, Stack Overflow publishes an annual Developer Survey to provide key insights into the developer community, as well as other surveys from time to time, some of which include demographic (e.g., gender, race, age, etc.) and certain personal information. Your participation in the Developer Survey as well as any other Stack Overflow survey is entirely voluntary. In order to provide meaningful insights into the developer community, Stack Overflow may collect sensitive data about you, including, but not limited to your gender, race, sexual orientation, immigration status, and your location. We consider this data to be among our most sensitive and have taken elevated measures to protect such data as well as to de-identify your name and other personal information from such data. We restrict such data within Stack Overflow on a privileged basis to only those that need to see it and

while we may share your personal information with our third party providers for our research, we do so only after imposing contractual obligations of confidentiality on such third parties and only where necessary for limited research purposes. When we publish our surveys, they are published as a compilation of anonymized data. We may ask for your consent to contact you following completion of the surveys. If you have any questions concerning the Developer Survey or other survey and our security measures to protect your sensitive personal data, please contact us at: privacy@stackoverflow.com. Advertising on our Network We work with third party advertisers to deliver ads on our Network. This involves tracking, targeting and measuring, using your personal data. We will get consent through our online consent management portal, or via other consent mechanisms before we can use your personal data in this way.

Targeting: Advertisers can target their campaigns using topic tags. Advertisers can target by geography, which is identified through Google Ad Manager, based on a visitor's IP address (which is truncated by Google). Advertisers can target by industry or company size, which is identified using IP address and matching that to Dunn Bradstreet data. Measuring: We allow advertisers to use MOAT, Doubleverify, and IAS tags on their campaigns to monitor for fraud. We also offer conversion pixels so advertisers can track conversions that happen within 30 days of an ad being served. Other Third Party Online and Digital Advertising When you access our websites, complete online contact forms, or interact with us, we may use your information to provide you with advertising content.

Google Ads We use Google Ads to display personalized ads to users. You can find out more or opt out by visiting Google's help center. We will only process your data for any online or digital advertising on the basis of your informed consent, which can be withdrawn at any time, or if we have a legitimate interest. Your data can only be used by Stack Overflow and will not be shared with other parties.

Microsoft Advertising We use Microsoft services (e.g., Bing) for personalized advertising purposes, including Match lists, user event tracking (UET) and retargeting. When we do this, Microsoft collects or receives Personal Data from us to provide Microsoft advertising. Your data can only be used by Stack Overflow and will not be shared with other parties. Where applicable, (for example, if you reside in the EEA or where EEA law applies), we will get your consent before we can share your personal data with Microsoft. You can find out more by visiting Microsoft Privacy Statements. You can also control the

use of your data for interest-based advertising from Microsoft by visiting their opt-out page.

Employer Branding Stack Overflow provides services that enable organizations to provide branded content on our website. We maintain a dedicated page on the public Stack Overflow page that enables our customers to set up a page and create branded content. The Customer has an admin key to the page, controls the content and can invite others to be admins. Most content is static and just resides on the page. Dynamic content (posting updates, giving advice, etc.) lives externally to the page and can be accessed by users through links provided by our customer. Public users can "follow" the page through getting a weekly email digest and they can unfollow or opt out of receiving emails from the Customer. The Customer has no direct interaction with the public website users on the page. Users must ensure that they read any privacy notices provided by customers on the page as well as this privacy notice.

WHO DO WE SHARE YOUR PERSONAL DATA WITH?

We share personal data with: Companies within the Stack Exchange network. Third parties who provide professional services, including but not limited to accountants; banking, insurance and insurance broking services for us. Third parties we use to help deliver our products and services to you, e.g., payment service providers, payment processors, warehouses and delivery companies; cloud service providers, e.g., Microsoft Azure, service providers that help us carry out certain tasks, including order fulfilment, customer service providers, maintaining technology and related infrastructure, serving and targeting ads, measuring performance, managing and analysing research, email distribution, managing marketing e.g., Marketo and promotions and surveys. Additionally, to improve user experience, we offer single sign-on solutions for account login and these third parties (including Facebook and Google) may receive information from these services when you elect to use them. When Stack Overflow shares your personal information and other collected information with third party service providers, we require that they use your information only for the purpose of providing services to us and consistent with this privacy policy. Third parties that we partner with to deliver products and services, including certain marketing and features to you, such as ; and Collectives on Stack Overflow (please read the). Employer Branding Collectives Privacy Notice Advertising partners – When you visit or interact with our websites, e.g., when you use the Public

Network, we and authorized third parties will, with your permission, place certain cookies on your device and your account activity as described in more detail within our . We use this information to serve you certain advertising content. We also partner with other third parties, such as Google Ads and Microsoft Bing, to serve advertising content and manage advertising campaigns. When we use Google Ads or Microsoft Bing Customer Match for advertising campaigns, your personal data will be protected using hashed codes. Google users can control the ads that they see on Google services, including Customer Match ads, in their Google Ads Settings. More information on how to control your account settings or to opt out of direct marketing campaigns can be found under section. You can also opt out in your . cookie policy Your Choices Profile Settings Event sponsors and partners – we may share your personal data with sponsors of Stack Overflow events and partners whom we hold events with for marketing purposes when you have given your permission for us to do so. Other third parties approved by you, e.g., third parties that run our Developer Survey, social media sites you choose to link your account to or third party payment providers. We may make available APIs to enable users to download publicly available information from our websites. Authorities who require reporting of processing activities in certain circumstances, for example, for anti-money laundering, sanctions and other regulatory obligations. We will not disclose your personal information to legal authorities except where such disclosure is by lawful request, including to meet legitimate national security or law enforcement demands (including a subpoena, court order, or other lawful legal demand by a legal authority with lawful jurisdiction) or in the public interest in compliance with applicable laws. In some cases, we may also release your personal information to defend our legal interests. If we choose to sell, transfer, or merge parts of our business or our assets, your personal data would be shared with such third parties as part of such a transaction. Alternatively, we may seek to acquire other businesses or merge with them. If a change happens to our business, then the new owners may use your personal data in the same way as set out in this privacy notice. We require all third parties to respect the security of your personal data and to treat it in accordance with the law. We do not allow our third-party service providers to use your personal data for their own purposes and only permit them to process your personal data for specified purposes and in accordance with contractual obligations which ensure adequate protection for your personal data.

HOW LONG DO WE RETAIN YOUR PERSONAL DATA?

We will keep your personal data only for as long as is necessary for the purposes set out in this privacy notice and to fulfil our legal obligations. We will not keep more data than we need. While you remain a user of the Public Network or a customer, we will retain your data in order to keep providing services to you. We are required to keep some basic information about our customers for a longer period in accordance with applicable tax and other laws. In some circumstances, you can ask us to delete your data. Your choices - how to access and control your personal data You can make choices about the collection and use of your data by Stack Overflow in the various ways set out below. You may change or correct your account settings and choose not to share certain information by creating a pseudonymous account. You may also remove certain optional information that you no longer wish to be publicly displayed, such as your photo, or job history.

You can also control your personal data that Stack Overflow has collected and exercise your data protection rights as set out in the section below, or by using the tools described below. Please be aware that in some cases, your ability to access or control your personal data will be limited under applicable law. Promotional emails, SMS messages, telephone calls and postal mail from Stack Overflow

When we send promotional emails and texts, we will give you an 'unsubscribe' option to opt out of receiving such messages. You can opt out of receiving telephone calls. Occasionally, we may send promotional mail, you can let us know if you do not wish to receive mail by contacting us. You can also opt out in your Profile Settings. Interest based advertising We may carry out interest-based advertising through third party services such as Microsoft or Google, as described under Advertising, above. You can opt out of such advertising in the following ways: You can control the use of your data for interest-based advertising from Microsoft by visiting their opt-out page. You can opt out of advertising from Google or find out more by visiting Google's help center.

Appendix C: GitHub General Privacy Statement

This appendix includes GitHub's privacy policy statement [46].

This is where we describe how we handle your "Personal Data", which is information that is directly linked or can be linked to you. It applies to the Personal Data that GitHub, Inc. or GitHub B.V., processes as the "Data Controller" when you interact with websites, applications, and services that display this Statement (collectively, "Services"). This Statement does not apply to services or products that do not display this Statement, such as Previews, where relevant. End User Notice: Organization-Provided GitHub Accounts When a school or employer supplies your GitHub account, they assume the role of Data Controller for most Personal Data used in our Services. This enables them to: Manage and administer your GitHub account, including adjusting privacy settings. Access and utilize your Personal Data, which includes details on how you use the Services, as well as your content and files. Should you access a GitHub Service through an account provided by an organization, such as your employer or school, the organization becomes the Data Controller, and this Privacy Statement's direct applicability to you changes. Even so, GitHub remains dedicated to preserving your privacy rights. In such circumstances, GitHub functions as a Data Processor, adhering to the Data Controller's instructions regarding your Personal Data's processing. A Data Protection Agreement governs the relationship between GitHub and the Data Controller. For further details regarding their privacy practices, please refer to the privacy statement of the organization providing your account. In cases where your organization grants access to GitHub products, GitHub acts as the Data Controller solely for specific processing activities. These activities are clearly defined in a contractual agreement with your organization, known as a Data Protection

Agreement. You can review our standard Data Protection Agreement at GitHub Data Protection Agreement. For those limited purposes, this Statement governs the handling of your Personal Data. For all other aspects of GitHub product usage, your organization's policies apply.

Third Party Access and Data Protection When you use third-party extensions, integrations, or follow references and links within our Services, the privacy policies of these third parties apply to any Personal Data you provide or consent to share with them. Their privacy statements will govern how this data is processed.

Personal Data We Collect Personal Data is collected from you directly, automatically from your device, and also from third parties. The Personal Data GitHub processes when you use the Services depends on variables like how you interact with our Services (such as through web interfaces, desktop or mobile applications), the features you use (such as pull requests, Codespaces, or GitHub Copilot) and your method of accessing the Services (your preferred IDE). Below, we detail the information we collect through each of these channels: From You Account Data: We collect certain information when you open an account such as your GitHub handle, name, email address, password, payment information and transaction information. User Content and Files: When you use our Services, we collect Personal Data included as part of the information you provide such as code, inputs, text, documents, images, or feedback. Demographic information: In some cases, you provide us with ethnicity, gender, or similar demographic details. Feedback Data: This consists of information you submit through surveys, reviews, or interactive features. Payment Information: For paid subscriptions, we collect details like name, billing address, and payment specifics. Profile Information: We collect information to create a user profile, which may include a photo, additional email addresses, job title, or biography. Sales and Marketing Data: This includes information provided for promotional communications, such as name, email address, and company name. Support Data: When you seek customer support, we collect details like code, text, or multimedia files. Automatically Buttons, Tools, and Content from Other Companies: Our Services may contain links or buttons that lead to third-party services like Twitter or LinkedIn. Use of these features may result in data collection. Engaging with these buttons, tools, or content may automatically send certain browser information to these companies. Please review the

privacy statements of these companies for more information. Essential Cookies and Similar Tracking Technologies: We use cookies and similar technologies to provide essential functionality like storing settings and recognizing you while using our Services. Nonessential Cookies: Depending on your jurisdiction, we may use online analytics products that use cookies to help us analyze how de-identified users use our Services and to enhance your experience when you use the Services. We may also employ third-party Cookies to gather data for interestbased advertising. In some jurisdictions, we only use nonessential cookies after obtaining your consent. See this section for more details and control options.

Email Marketing Interactions: Our emails may have web beacons that offer information on your device type, email client, email reception, opens, and link clicks. Geolocation Information: Depending on the Service's functionality, we collect regional geolocation data Service Usage Information: We collect data about your interactions with the Services, such as IP address, device information, session details, date and time of requests, device type and ID, operating system and application version, information related to your contributions to repositories, and performance of specific features or Services. Website Usage Data: We automatically log data about your Website interactions, including the referring site, date and time of visit, pages viewed, and links clicked. From Third Parties Information from Other Users of the Services: Other users may share information about you when they submit issues and comments. We may also receive information about you if you are identified as a representative or administrator on your company's account. Publicly Available Sources: We may acquire information about you from publicly available sources like public GitHub repositories. Services you linked to your GitHub account: When you or your administrator integrate third-party apps or services with our Services, we receive information based on your settings with those services. This can include details like your name and email from services like Google for authentication. The information we receive depends on the third-party's settings and privacy policies. Always review these to understand what data is shared with our Services. Vendors, Partners, and Affiliates: We may receive information about you from third parties, like vendors, resellers, partners, or affiliates for the purposes outlined in this statement. Processing Purposes: How We Use Your Personal Data The Personal Data we process depends on your interaction and access methods with our Services, including the interfaces

(web, desktop, mobile apps), features used (pull requests, Codespaces, GitHub Copilot), and your preferred access tools (like your IDE). This section details all the potential ways GitHub may process your Personal Data: Business Operations: We use Personal Data for activities like billing, accounting, and compensation. This includes creating aggregated statistical data for internal reporting, financial reporting, revenue planning, capacity planning, and forecast modeling (including product strategy). Communication: We use Personal Data to inform you about new Services, features, offers, promotions, and other pertinent information. This also includes sending confirmations, invoices, technical notices, updates, security alerts, and administrative messages. Inference: We generate new information from other data we collect to derive likely preferences or other characteristics. For instance, we infer your general geographic location based on your IP address. Personalization: We use Personal Data to customize the Service to your preferences, to evaluate the effectiveness of enterprise business ads and promotional communications, and to ensure a seamless and consistent user experience. Safety and Security: To promote safety, integrity, and security across our Services, we process Personal Data, using both automated and, at times, manual techniques for abuse detection, prevention, and violations of terms of service. Service Provision: We use Personal Data to deliver and update our Services as configured and used by You, and to make ongoing personalized experiences and recommendations. Troubleshooting: We use Personal Data to identify and resolve technical issues. Ongoing Service Performance: Personal Data helps us keep the Services up to date and performant, and meet user productivity, reliability, efficacy, quality, privacy, accessibility and security needs. Complying with and resolving legal obligations: including responding to Data Subject Requests for Personal Data processed by GitHub as Controller (for example website data), tax requirements, agreements and disputes. Delivering Professional Services: We use Personal Data to deliver training, consulting or implementation ("Professional Services"). This includes providing technical support, professional planning, advice, guidance, data migration, deployment, and solution/software development services. Improving Professional Services: Enhancing delivery, efficacy, quality, and security of Professional Services and the underlying product(s) based on issues identified while providing Professional Services, including fixing software defects, and otherwise keeping the Professional Services up to date and performant. When carrying out these activities, GitHub practices data minimization and uses

the minimum amount of Personal Information required. Sharing of Personal Data We may share Personal Data with the following recipients: Abuse and Fraud Prevention Entities: We may disclose Personal Data based on a good faith belief it is needed to prevent fraud, abuse, or attacks on our Services, or to protect the safety of GitHub and our users. Affiliates: Personal Data may be shared with GitHub affiliates, including Microsoft, to facilitate customer service, marketing and advertising, order fulfillment, billing, technical support, and legal and compliance obligations. Our affiliates may only use the Personal Data in a manner consistent with this Privacy Statement. GitHub Organization Accounts: If an organization adds you to their GitHub account, we might share Personal Data with that organization to fulfill the commercial relationship. In such a case, your use of the Services is protected by a data protection agreement and terms between your organization and GitHub Competent Authorities: We may disclose Personal Data to authorized law enforcement, regulators, courts, or other public authorities in response to lawful requests or to protect our rights and safety. Please refer to our Guidelines for Legal Requests of User Data for more information. Corporate Transaction Entities: we might disclose Personal Data within the limits of the law and in accordance with this Privacy Statement for strategic business transactions such as sales or a merger. Partners and Resellers: We cooperate with third-parties that offer sales, consulting, support, and technical services for our Services. We may share your data with these partners and resellers where allowed, and with your consent when required. Subprocessors and Service Providers: We may use vendors to provide services on our behalf, including hosting, marketing, advertising, social, analytics, support ticketing, credit card processing, or security services. They are bound by contractual obligations to ensure the security, privacy, and confidentiality of your information. Visual Studio Code (GitHub Codespaces): GitHub Codespaces and github.dev offer Visual Studio Code in a web browser, where some telemetry is collected by default. Details on telemetry collection are on the VS Code website. Other Third-party Applications: Upon your instruction, we may share Personal Data with third-party applications available on our Marketplace. You are responsible for the data you instruct us to share with these applications. Other Users and the Public: Depending on your account settings, we may share Personal Data with other users of the Services and the public. You control what information is made public. To adjust your settings, visit User Settings in your profile. Please be aware that any information you share in a collaborative context

may become publicly accessible. Private repositories: GitHub Access If your GitHub account has private repositories, you control the access to that information. GitHub personnel does not access private repository information without your consent except as provided in this Privacy Statement and for: security purposes automated scanning or manual review for known vulnerabilities, active malware, or other content known to violate our Terms of Service to assist the repository owner with a support matter to maintain the integrity of the Services, or to comply with our legal obligations if we have reason to believe the contents are in violation of the law. GitHub will provide you with notice regarding private repository access unless doing so is prohibited by law or if GitHub acted in response to a security threat or other risk to security.