Social Network Analysis and Mining (2024) 14:16
https://doi.org/10.1007/s13278-023-01171-z

ORIGINAL ARTICLE

Cybersecurity discussions in Stack Overflow: a developer-centred analysis of engagement and self-disclosure behaviour

Nicolás E. Díaz Ferreyra[1] · Melina Vidoni[2] · Maritta Heisel[3] · Riccardo Scandariato[1]

Received: 28 February 2023 / Revised: 22 June 2023 / Accepted: 15 November 2023

Abstract
Stack Overflow (SO) is a popular platform among developers seeking advice on various software-related topics, including privacy and security. As for many knowledge-sharing websites, the value of SO depends largely on users' engagement, namely their willingness to answer, comment or post technical questions. Still, many of these questions (including cybersecurity-related ones) remain unanswered, putting the site's relevance and reputation into jeopardy. Hence, it is important to understand users' participation in privacy and security discussions to promote engagement and foster the exchange of such expertise. *Objective*: Based on prior findings on online social networks, this work elaborates on the interplay between users' engagement and their privacy practices in SO. Particularly, it analyses developers' self-disclosure behaviour regarding profile visibility and their involvement in discussions related to privacy and security. *Method*: We followed a mixed-methods approach by (i) analysing SO data from 1239 cybersecurity-tagged questions along with 7048 user profiles, and (ii) conducting an anonymous online survey (N=64). *Results*: About 33% of the questions we retrieved had no answer, whereas more than 50% had no accepted answer. We observed that *proactive* users tend to disclose significantly less information in their profiles than *reactive* and *unengaged* ones. However, no correlations were found between these engagement categories and privacy-related constructs such as *perceived control* or *general privacy concerns*. *Implications*: These findings contribute to (i) a better understanding of developers' engagement towards privacy and security topics, and (ii) to shape strategies promoting the exchange of cybersecurity expertise in SO.

Keywords Stack Overflow · Usable privacy and security · Engagement · Self-disclosure · R programming · Python · Social coding platforms

1 Introduction

The last decade has put privacy in the spotlight of software development, as new legal frameworks emerged to safeguard people's data protection rights and promote responsible engineering practices. One clear example is the EU General Data Protection Regulation (GDPR) (Parliament 2016) which has introduced strong legal provisions seeking to enforce software companies to comply with a set of privacy principles including transparency, fairness, and informed consent. More recently, as the software industry moves towards the development of Artificial Intelligence (AI) applications, a new regulatory framework is in sight (European Comission 2021), promising to strengthen the protection and governance of personal data in AI systems. In turn, companies and organisations have been urged to adopt privacy-by-design practices to comply with current regulations. Nevertheless, this has also raised questions and concerns among software developers on how to effectively translate these legal provisions and privacy principles into technical solutions (Sirur et al. 2018).

Question-Answer (Q&A) platforms are a valuable resource for both experienced and junior programmers

✉ Nicolás E. Díaz Ferreyra
nicolas.diaz-ferreyra@tuhh.de

Melina Vidoni
melina.vidoni@anu.edu.au

Maritta Heisel
maritta.heisel@uni-due.de

Riccardo Scandariato
riccardo.scandariato@tuhh.de

[1] Hamburg University of Technology, Hamburg, Germany

[2] Australian National University, Canberra, Australia

[3] University of Duisburg-Essen, Duisburg, Germany

Published online: 22 December 2023

seeking support in their software development tasks. Stack Overflow (SO)[1] is among the largest Q&A platforms in which developers participate in discussions related to performance issues, bugs, and code workarounds (Ahmed and Srivastava 2017). Given the increasing importance of cybersecurity in software engineering, a large number of questions regarding privacy, security, and data protection have been posited and addressed by SO users. Particularly, issues related to GDPR compliance, privacy policies, and access-control are some of the most popular privacy-related discussions in SO (Tahaei et al. 2020; Lopez et al. 2018). Still, privacy and security-related topics receive little attention in comparison with others such as data science, big data, and mobile operating systems.[2] Albeit this suggests a low engagement towards cybersecurity discussions within the SO community, it also reveals an overall tendency among software developers to overlook privacy and security aspects of their code (Senarath and Arachchilage 2018; Assal and Chiasson 2018; Hadar et al. 2018).

1.1 Motivation

Developers play a key role in embedding privacy and security principles into the core architecture of information systems (Hadar et al. 2018). However, many often fail to create secure software solutions that successfully preserve users' privacy and data protection rights (Senarath and Arachchilage 2018; Hadar et al. 2018). Over the last years, a growing body of research has leveraged the SO's dataset to identify and characterise cybersecurity trends among software practitioners. Prior work has investigated developers' motivations (Lopez et al. 2018), knowledge gaps (Tahaei et al. 2020), and concerns towards privacy and security (Lopez et al. 2019). However, "*answer-hungry*" questions are still a common phenomenon and an ongoing issue within Q&A websites (i.e. questions remaining unanswered or unresolved) (Gao et al. 2020). Being SO a community frequented by more than 100 Million developers per month,[3] users' commitment towards timely and high-quality answers becomes critical for the platform's reputation and success. Former research has sought to understand users' motivations (and amotivations) when it comes to participation in Q&A forums (Yang et al. 2014; Chua and Banerjee 2015; Adaji and Vassileva 2016). Yet, little effort has been made to characterise users' engagement in cybersecurity discussions in SO. That is, on providing evidence and actionable information about community members participating actively (or not) in such exchanges.

Individuals' engagement in Online Social Networks (OSNs) like Facebook has been extensively investigated from the perspective of privacy concerns. Such research has analysed the connection between users' self-disclosure decisions (e.g. the amount of private information they reveal inside profiles and posts) and their engagement in these platforms (e.g. number and quality of OSN posts) (Kayes et al. 2015; Choi and Sung 2018; Staddon et al. 2012). Overall, such research has not only contributed to a better understanding of users' privacy concerns and practices but has also paved the road for the development of user-centred technologies. That is, for the elaboration of methods and tools aiming to support and guide users' interaction in OSN environments (Seamons 2022). However, to the extent of our knowledge, the role of privacy-related behaviour has not been closely investigated within Q&A platforms like SO. Particularly, the interplay between developers' self-disclosure practices and their engagement in discussion threads has not been yet explored under the lens of privacy and security benchmarks.

1.2 Contribution and research questions

SO is a valuable resource for developers seeking advice about multiple aspects of software development. Given the increasing importance of cybersecurity in software engineering, it becomes necessary to foster the engagement among its users towards privacy and security-related discussions. Hence, this work aims at contributing to ongoing research in SO by investigating the interplay between users' self-disclosure decisions and their engagement in cybersecurity discussions. All in all, the research questions (RQs) this paper seeks to answer are:

- **RQ1**: *Are users' self-disclosure behaviour associated with their engagement in cybersecurity discussions?* Prior studies in OSNs (in general) and Q&A platforms (in particular) have shown correlations between users' engagement and self-disclosure practices (e.g. Adaji and Vassileva 2016; Kayes et al. 2015; Vargo and Matsubara 2018). Hence, this RQ aims at zooming into developers' decisions regarding profile visibility and their participation in discussions about privacy and security. Particularly, it seeks to investigate whether different self-disclosure patterns exist across SO users who involve themselves actively in such discussions, and those who do not.
- **RQ2**: *Are privacy-related constructs associated with users' engagement in cybersecurity discussions?* As with RQ1, former studies have delved into the relation between psychological constructs (e.g. perceived risks

[1] https://stackoverflow.com

[2] By May 2021, the amount of *security*- and *privacy*-related questions was around 53.000, whereas for *Android* and *iOS* it was over 1.900.000 https://stackoverflow.com .

[3] https://stackoverflow.co/advertising/audience/

and control) and peoples' engagement within OSNs (e.g. Staddon et al. 2012; Jozani et al. 2020). The purpose of this RQ is to examine whether such correlations also take place in SO but regarding users' participation in discussions about privacy and security.

To answer these RQs, we have followed a mixed-method approach combining the analysis of data collected from an online survey and information retrieved from SO user profiles. The results of our analysis show significant differences in the self-disclosure practices (i.e. with regard to profile visibility) of users contributing actively to discussions about data protection and information security, and those who do not. These findings not only contribute to a better understanding of users' engagement in such discussions, but also to solutions addressing "answer-hungry" questions in Q&A platforms. Particularly, for the elaboration of incentive strategies and recommender systems promoting the exchange of cybersecurity expertise in SO.

Paper Structure. Sect. 2 discusses related work and gives and overview of the paper's theoretical background. Section 3 describes the methodology employed for the study in terms of data collection, aggregation, and survey design. Section 4 reports the results of our analysis, and Sect. 5 discusses them. Section 6 summarises limitations and threats to validity. Section 7 concludes this work.

2 Background and related work

A growing amount of literature has zoomed into cybersecurity discussions in SO and engagement patterns in OSNs. This section summarises related work elaborating on privacy and security insights gathered through SO. Alongside, we discuss research addressing privacy concerns as a rationale for users' engagement and self-disclosure behaviour in OSNs.

2.1 Cybersecurity discussions in SO

Given the Q&A affordances available within SO, this platform has been widely used as a proxy for understanding the cybersecurity concerns and practices of software engineers (Lopez et al. 2018; Tahaei et al. 2020; Lopez et al. 2019; Fischer et al. 2017). For instance, Lopez et al. (2018) conducted a qualitative analysis of SO discussion threads to understand the type of security support developers seek and provide online. Their findings suggest that security-related discussions in SO are rich in terms of technical help but also regarding developers' personal values and attitudes such as trust, fear, and sense of responsibility. In a follow-up article (Lopez et al. 2019), the authors gathered further insights on how security knowledge is built and fostered within the SO community. Overall, their results show that developers often tend towards security-related discussions within the context of technical solutions provided by others. In line with this, Tahaei et al. (2020) applied natural language processing techniques to unveil topics emerging within privacy-related questions. The outcome of such an analysis showed that privacy policies, access-control, and encryption are among the main privacy topics addressed by SO members. Moreover, the results of a follow-up study (Tahaei et al. 2022) indicate that privacy advice mostly relates to compliance and confidentiality issues.

At its core, SO is a peer-production community where knowledge is built from the interaction between developers seeking to clarify each other's technical inquires (Sengupta and Haythornthwaite 2020). Hence, users' participation and engagement are of utmost importance for the sustained development of the platform and the expertise crafted within it. Moreover, timely answers to questions are critical to the platform's efficiency and, thus, to its popularity. Nonetheless, prior research has systematically reported that many questions in SO receive little attention or even remain unanswered/unresolved (up to 30% by May 2022[4]). As a catalyst for developers' technical concerns and best practices, it is essential to understand the factors contributing to or impairing users' participation in SO. Prior work has tried to explain why some questions remain unanswered and even proposed machine learning models for predicting whether specific questions will be addressed or not (Ahmad et al. 2018). Still, the low engagement and the lack of answers to specific questions (including privacy and security-related ones) remain open issues (Gao et al. 2020). Hence, there is a call for empirical evidence to (i) help characterise users' engagement in cybersecurity discussions and (ii) elaborate strategies for boosting their participation in such discussions.

2.2 Insights from online social networks

Factors influencing people's participation in OSNs have been thoroughly investigated through the lens of privacy concerns. Moreover, prior work has closely analysed users' privacy practices, often accounting for correlations between OSN engagement and self-disclosure behaviour. Staddon et al. (2012), for instance, observed strong associations between privacy concerns and users' engagement on Facebook using an online survey. Their findings revealed that individuals expressing concerns about their privacy also report spending less time on the platform and sharing less content. Hence, they concluded that privacy concerns might play a significant role in people's engagement in OSNs. In line with this, a study by Choi and Sung (2018) showed that

[4] https://stackexchange.com/sites

privacy concerns are closely associated with active Instagram use (e.g. sharing content and interacting more with others) and people's selection of a particular OSN platform over others (e.g. Instagram over Snapchat). Alongside, research has systematically reported evidence on the so-called "privacy paradox", showing offsets between users' concerns and engagement in OSNs (Krämer and Schäwel 2020). Such evidence suggests that, despite expressing privacy concerns, people still join OSNs and disclose significant amounts of personal information.

When it comes to engagement in Q&A platforms, Kayes et al. (2015) investigated the interplay between users' privacy concerns and their participation in Yahoo! Answers. By considering changes in profile visibility as manifestations of privacy concerns, the authors unveiled correlations between users' self-disclosure behaviour and their platform contributions. Overall, they observed that users with a private profile contribute more often and with better content to the platform than those with a public one. Such findings can contribute substantially to the elaboration of Q&A recommendation approaches. For instance, one could leverage profile visibility for rooting unresolved questions to those users who are more likely to answer them (Kayes et al. 2015). Surprisingly, concerns and practices alike have not been thoroughly investigated in SO despite its Q&A and social network affordances. Moreover, to the extent of our knowledge, the relationship between engagement in cybersecurity topics and self-disclosure practices has not been yet explored nor investigated from a developer-centred perspective.

3 Methodology

We conducted a two-stage empirical study to identify nuances in the self-disclosure practices of users participating actively in cybersecurity discussions, and those who do not. For this, we created a dataset from 7048 SO profiles corresponding to *engaged* and *unengaged* users during the first stage of the study. This dataset was then leveraged on the second stage to conduct an anonymous online survey. Both experimental stages are described in detail in the following subsections.

3.1 Data collection

To identify users concerned with cybersecurity topics, we first conducted an analysis of privacy and security-related conversations in SO. Such an analysis consisted in the identification of cybersecurity-relevant conversation threads through their corresponding user-assigned tags. For this, we used SO's Tag Explorer[5] for the definition of tag sets which were used thereafter to mine relevant conversations. Particularly a set of *topic tags* plus two *language tags* were employed in the identification of cybersecurity-relevant discussions.

We included `privacy`,`security`, `privacy-policy`, `code-access-security`, `data-security`, `network-security`, and `gdpr-consentform` as topic tags.[6] Additionally, `r` and `python` were used as *language tags* given the increasing popularity of these languages within the data science community (Moutidis and Williams 2021). Thereby, we sought to narrow down the scope of the study mainly to data science practitioners as they are prone to handle sensitive data (e.g. medical records, biometric data, demographics). Furthermore, their cybersecurity practices can have a great impact on automated decision-making systems (e.g. biases, discrimination).

3.1.1 Discussions dataset (D1)

Each *topic tag* was explored in combination with each *language tag*, resulting in 14 tag searches. To maximise the size of the dataset, we did not include additional restrictions such as time of posting, the existence of an approved answer, upvotes, or downvotes. Both search and extraction were executed through an R-based mining package included in the StackExchange API.[7] We conducted fourteen independent searches (i.e. one per tag combination) using the `search/advanced` endpoint and a `tag filter` provided by the API itself. **By the end of the mining process, a total of 1239 questions/posts were retrieved from SO (Figure 1).**

Questions posted in SO can be answered or commented on by other platform members. The main difference is that the latter asks for clarification instead of describing a suitable solution. One question can trigger several answers and comments (to the main question or to others' answers) from other SO users interested in the discussion topic. Therefore, such comments and answers are also relevant for identifying SO profiles corresponding to individuals who engage in cybersecurity discussions. Consequently, answers and comments associated with each of the 1239 questions were also mined and included in a *discussions dataset* D_1. After this additional mining process, $\boldsymbol{D_1}$ contained 1239 questions, 2558 comments to questions, 1811 answers, and 2373 comments to answers.

[5] See: https://stackoverflow.com/tags

[6] It is worth mentioning that, at the moment of conducting this study, these were the only cybersecurity-related tags available in SO.

[7] See: https://api.stackexchange.com/

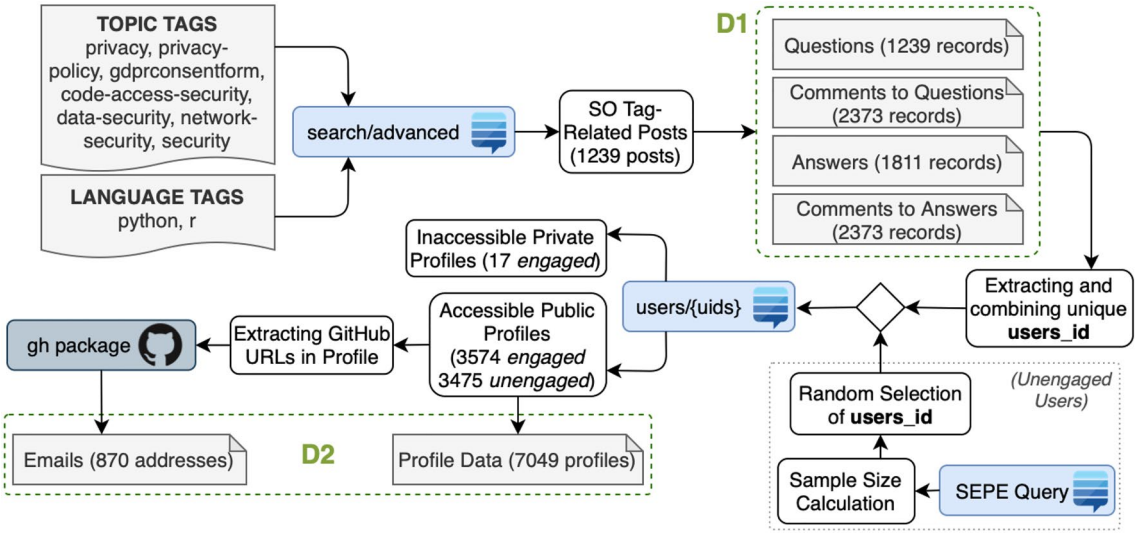

Fig. 1 Mining process followed to extract and generate both datasets

3.1.2 Profiles dataset (D2)

i. *Engaged user profiles:* The information contained in D_1 allowed us to identify the SO ids of those users who have either posted a question, provided an answer, or posted a comment deemed as cybersecurity-relevant. Overall, 3591 unique ids were retrieved, from which only 17 corresponded to users with fully private SO profiles. The remaining **3574 ids** were used to mine the public information disclosed in their profiles through the StackExchange API (i.e. via the `users/{ids}` endpoint). The email address of some of them was also mined using the GitHub (GH) URL available in the profiles (email addresses are never included in SO profile pages). This step was necessary to recruit participants afterwards for the online survey. This complementary mining process was executed using the R package `gh`[8] resulting in 457 unique e-mail addresses corresponding to *engaged* users. Such information was included in the profiles dataset D_2 along with the rest of the profile information extracted from SO.

ii. *Unengaged user profiles:* In order to populate D_2 with profile information from *unengaged* users, we first estimated a representative sample size for such a subgroup. For this, we run a query to determine how many users have participated on each language tag[9] using Stack Exchange's Data Explorer (SEDE). The result of this query gave 46038 users for the `r` tag, and 777587 for `python` tag. Next, we mined the profile information from a representative sample of these two groups with a 99% confidence and a margin of error of 3%. Such information was mined directly from the `users/{uids}` endpoint, ensuring that the corresponding SO ids were not already part of the *engaged* group, and were not repeated across each language. Overall, we obtained 1830 Python users and 1645 R users (**3475 in total**). These results were merged into the D_2 dataset, using an additional variable to indicate whether this information corresponds to *engaged* or *unengaged* users. Like with the *engaged* profiles, we collected the e-mail addresses of 413 *unengaged* users via GH (Fig. 1).

3.2 Data aggregation

We parsed the information collected in both datasets to compute two variables of interest: (i) the amount of information users disclose in their profiles, and (ii) their engagement in cybersecurity discussions. The following subsections describe these variables plus an additional analysis we conducted to understand self-disclosure through display names.

3.2.1 Amount of self-disclosure

SO allows users to include the following information in their profiles: *display name* (with a maximum of 30 characters), *location* (as a text field), *title* (available in the profile, but merged into the display name when using the API), *about me* (HTML-friendly text box of up to 3000 characters), a *website* link, links to *Twitter* and *GitHub* profiles, and a *profile*

[8] See: https://cloud.r-project.org/web/packages/gh/index.html

[9] Query: https://data.stackexchange.com/stackoverflow/query/1392147

picture (if not used, the system assigns a randomised avatar). To compute a metric reflecting the amount of personal information revealed in a profile, we assigned a normalised variable (i.e. ranging from 0 to 1) to each field except for the title. The value for each particular variable was estimated as follows:

- We gave each link (*website*, *Twitter* and *GitHub*) a value of 1 if it was filled in the user's profile, and 0 if not.
- The *location* variable was calculated as the links (i.e. 1 if it was completed and 0 if not). Since users can obfuscate this field (e.g. by using nicknames or aliases), we conducted a card sorting analysis to estimate the reliability of this coding schema. From this analysis, we concluded that location information could be considered accurate if present.
- The variable corresponding to the *display name* was computed as the proportion of used characters over the total available (30 characters). As with *location*, we completed another card sorting analysis to obtain further reliability insights. Once again, we concluded that the information present in this field could be considered accurate. Both card-sorting analyses can be found in the Appendix A.
- The *profile image* was retrieved as an URL address during the data collection process. To determine whether an image corresponds to a *custom* or a *default* one we compared its URL against a collection of Gravatar[10] URLs (Gravatar pictures are frequently used as default in SO profiles). Using regular expressions, we assigned a 0 value to those profile pictures found in the Gravatar database. Otherwise, they were considered as *custom* and given a value of 1.
- The *about me* field can have up to 3000 characters allowing HTML formatting. The HTML tags were removed through an R script, and the proportion of used characters was calculated to determine the corresponding disclosure value of this field. This approach assumes that, as more characters are included, more personal information is being revealed.

These normalised variables were aggregated into another variable named soProfDisclosure quantifying the amount of personal information disclosed in a SO profile:

$$\text{soProfDisclosure} = \frac{\text{attsVisibleInProfile}}{\text{maxAmountOfDisclosableAtts}}$$

where maxAmountOfDisclosableAtts corresponds to the maximum number of disclosable attribute values (7 in total), and attsVisibleInProfile to the summation of each normalised variable.

[10] See: https://en.gravatar.com/

3.2.2 Engagement in cybersecurity discussions

We classified users into engaged or unengaged, given their participation by computing the number of cybersecurity-relevant questions a user has posted ($\#Q$), the number of answers provided to such questions ($\#A$), and of corresponding comments. This last one was divided into comments to cybersecurity questions ($\#C_Q$) and comments to cybersecurity answers ($\#C_A$). Overall, if the sum $\#Q + \#A + \#C_Q + \#C_A$ was greater than 0, then the user was classified as *engaged* and, otherwise, as *unengaged*.

Also, we classified engaged users into *proactive* and *reactive* according to their tendency towards starting new discussion threads. Particularly, we considered *proactive* users those who place more questions than comments and answers. That is, in cases where $\#Q \geq \#A + \#C_Q + \#C_A$. Conversely, users posting more comments and answers than cybersecurity questions were classified as *reactive*. That is, when $\#Q < \#A + \#C_Q + \#C_A$.

3.3 Survey structure

To complement the analysis of profile information and discussion threads, we conducted an online survey within a subgroup of SO users. In particular, we aimed at measuring psychological constructs and antecedents to better understand developers' concerns and behaviour regarding cybersecurity. The questionnaire consisted of an introductory part and two main sections:

i. The **introductory section** provided information about the aim of the study along with the conditions for participation/withdrawal (participation was voluntary, and people were given a chance to withdraw at any time). We also included the contact details of the authors in case of further questions and enquiries.
ii. After accepting the survey's terms and conditions, participants were forwarded to the **first part** of the questionnaire. This part included questions eliciting demographic information (e.g. participants' gender, education level, and current work status) along with their prior experience in software development (e.g. years working with R or Python).
iii. The **second part** included a set of questions measuring the following constructs: *general privacy concerns* (GPC), *privacy concerns on social threats* (PCS), *privacy concerns on organisational threats* (PCO), *perceived privacy risk* (RSK), *perceived control* (PC), and *self-disclosure* (SD). We used well-established constructs and scales previously elaborated and validated by other authors (i.e. GPC by Buchanan et al. (2007) and the rest by Krasnova et al. (2009)). All questions were close-ended and measured using a

Table 1 Survey Self-Reported Demographic Data

Demographic	Ranges	Freq.	Resp. (%)
Gender	Female	1	1.56
	Male	61	95.31
	Non-Binary	1	1.56
	Prefer not to say	1	1.56
Educational level	Graduate Degree (MSc, PhD)	36	56.25
	High School or Less	3	4.69
	Some College	11	17.19
	Undergrad Degree (BSc, BA)	14	21.88
Employment status	Currently in School	1	1.56
	Currently in University	5	7.81
	Unemployed, not looking for work	2	3.13
	Unemployed, looking for work	1	1.56
	Working full-time	49	76.56
	Working part-time	6	9.38
Programming experience (R/Python)	<2 years	2	3.13
	2–5 years	14	21.88
	5–10 years	22	34.38
	>10 years	26	40.63
Other Programming Experience	2–5 years	3	4.69
	5–10 years	12	18.75
	>10 years	49	76.56

6-Point Likert scale to increase the responses' reliability. We also included an attention question by the end of this section to identify careless respondents and preserve the quality of the results (Kung et al. 2018).

3.3.1 Population and sampling

The survey was distributed through Qualtrics in April/May 2021 using the 870 email addresses collected during the mining process (Section 3.1.2). We gathered 69 responses, out of which five were filtered through the "attention control" question. The remaining 64 responses were considered for the corresponding analysis. Table 1 provides a detailed description of the study sample.

3.3.2 Ethical considerations

The methodology used in this paper was approved by the Australian National University Human Ethics Research Committee (HREC) with project code 2021-24127, and conducted in accordance with the Declaration of Helsinki. Participants received information about the study procedure (including data privacy statements) and were asked for their informed consent. They also had the chance to withdraw at anytime without their answers being recorded. All survey protocols, responses, and data collected for this study are available in the paper's **Replication Package**.[11]

Table 2 Question status indicators

Indicator	Frequency	Total (%)
Has answers	825	67
Has accepted answers	588	47
Has score > 0	719	58
Has comments	489	39
Closed	94	8

4 Results

We conducted several statistical analyses over the information collected from SO and the responses obtained through the online survey. We conducted a *t*-Test (Ross and Willson 2017a) followed by an ANOVA test (Ross and Willson 2017b) to identify significant differences in the self-disclosure practices of engaged and unengaged users. The results of these tests were complemented afterwards with an analysis of the survey data.

4.1 Privacy and security discussions (SO Q&A data)

A total of 1239 cybersecurity-related questions were collected from SO using the StackExchange API (as explained in Sect. 3.1.1). As shown in Table 2, around 67% of these questions had at least one answer (*answered*), and about 47% received an answer considered adequate by the user who asked the question (*accepted*). Another 58% had a positive score (i.e. a positive difference between up-votes and down-votes), whereas 39% of the questions received at least one comment. SO also allows experienced community members to close questions that are either off-topic or may need further clarification. We observe that around 8% of the questions in our dataset fall into this category.

4.2 Self-disclosure practices (SO profile data)

As mentioned in Sect. 3.1.2, from the 7049 profiles retrieved from SO, 3574 correspond to *engaged* users and 3475 to *unengaged* ones. Figure 2 illustrates the disclosure frequency of each profile attribute in our sample for each group of users. We can see that such frequencies are quite even across all attributes for both groups and that "display name" is an attribute everyone discloses. We can also observe that

[11] Available at: https://tinyurl.com/SO-CYBERSEC

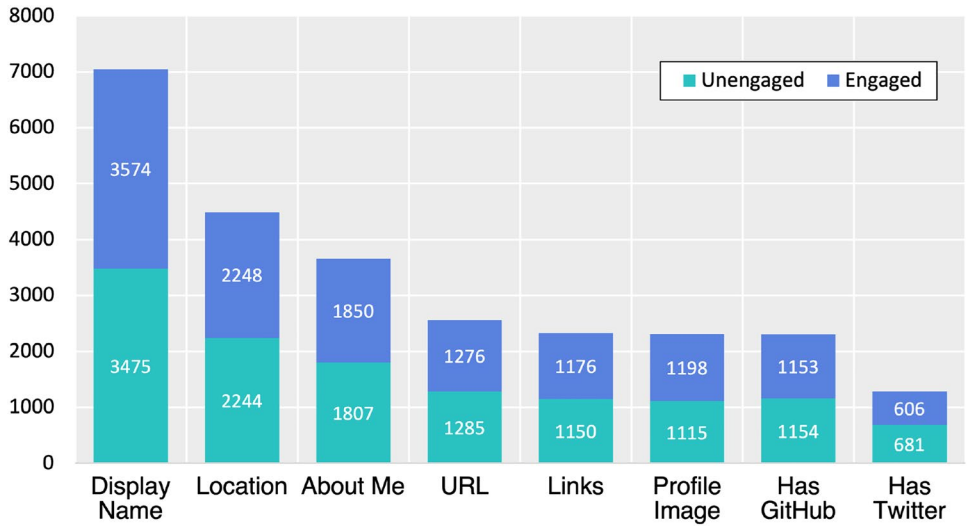

Fig. 2 Profile attributes disclosed by *unengaged* and *engaged* users (frequencies)

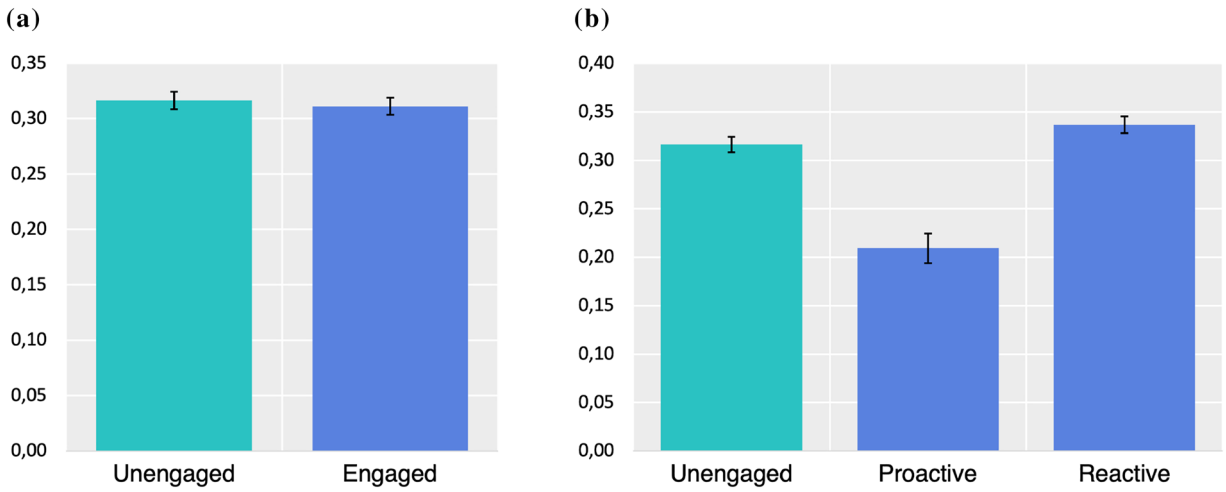

Fig. 3 Average self-disclosure of (a) *unengaged* and *engaged* users, and (b) *unengaged*, *proactive*, and *reactive* users

"location" and "about me" are among the most revealed profile attributes, whereas "has Twitter" is the least frequent one.

We ran an independent samples *t*-Test to identify significant differences in the amount of profile information disclosed by *engaged* and *unengaged* users. Since Levene's test for equality of variances resulted significant ($F_{1,7047} = 6.605$, $p = 0.10$), the corresponding t statistic was computed without assuming homogeneity of variances (Garson 2012). Overall, we found no significant differences in the average amount of self-disclosure between *engaged* and *unengaged* users ($t_{7027.424} = 0.918$, $p > 0.05$). This can be observed in Fig. 3-a. Hence, we conducted a follow-up ANOVA test to determine whether such differences exist among *unengaged*, *proactive*, and *reactive* users.

From the 3574 concerned profiles, 716 corresponded to *reactive* users and 2858 to *proactive* ones. In principle, we can observe differences in the amount of profile information disclosed across these 3 groups (Fig. 3-b). After conducting the ANOVA test (Table 5), we could confirm

Table 3 Games-Howell Test for Differences of Means

Diff. Levels	Diff. Means	SE	*p*	95% CI
Unengaged—Proactive	0.107*	0.008	0.000	(0.086, 0.127)
Unengaged—Reactive	−0.020*	0.006	0.002	(−0.034, −0.006)
Proactive—Unengaged	−0.107*	0.009	0.000	(−0.127, −0.086)
Proactive—Reactive	−0.127*	0.008	0.000	(−0.148, −0.106)
Reactive—Unengaged	0.020*	0.006	0.002	(0.006, 0.034)
Reactive—Proactive	0.127*	0.009	0.000	(0.106, 0.148)

* The mean difference is significant for $\alpha = 5\%$

Table 4 Multinomial Logistic Regression (estimates)

Group		B	SE	Sig.	Exp(B)
Proactive	Intercept	−0.991	0.062	0.000	
	% self-disclosure	−0.023	0.002	0.000	0.978
Reactive	Intercept	−0.313	0.043	0.000	
	% self-disclosure	0.004	0.001	0.001	1.004

that such differences were indeed statistically significant ($F_{2,7046} = 86.180,\ p < 0.05,\ \eta^2 = 0.024$). To determine where these differences actually occur, we ran an additional non-parametric posthoc test. We chose a Games-Howell test since Levene's statistic suggested no equal variances within the sample ($F_{2,7046} = 31.772,\ p < 0.05$). This analysis revealed significant differences ($p < 0.05$) in the average amount of self-disclosure across all paired groups (Table 3). That is, between *unengaged-proactive* (−0.107), *unengaged-reactive* (−0.020), and *proactive-reactive* (−0.127).

Finally, we conducted a multinomial logistic regression (Garson 2014) to obtain further insights on the self-disclosure practices of SO users. For this, we considered the unengaged users as the baseline category against which the other groups (i.e. proactive and reactive) should be compared. The parameter estimates of the resulting model are summarised in Table 4. As it can be observed, the percentage of information disclosed in a profile (*% self-disclosure*) is a significant predictor for both proactive and reactive user categories ($p < 0.05$).

On the one hand, for every one-unit increase on *%self-disclosure*, the likelihood a user has of falling in the proactive category decreases by 2.2% (i.e. relative to falling in the unengaged group). Conversely, such a likelihood increases by 0.4% for the reactive category. This model is a significant improvement in fit over an intercept model with no predictors ($\chi^2_2 = 181.388, p < 0.05$). However, it does not fit well to the data, which makes it not adequate for prediction purposes (Pearson's $\chi^2_{170} = 256.978, p < 0.05$).

4.3 Privacy-related constructs (survey data)

As shown in Table 1, 76.56% of the survey respondents worked full time and had more than 10 years of programming experience. Another 75% reported having more than 5 years of experience working with R or Python, and 56.25% having a graduate degree. In terms of gender, 61 out of the 64 participants were male, 1 was a woman, 1 non-binary, and 1 preferred not to reveal it.

Following the same user categories investigated in Sect. 4.2, we conducted a one-way ANOVA test to analyse the privacy-related constructs elicited in the second part of the survey (i.e. GPC, PCS, PCO, RSK, PC, and SD). From the 64 participants, 33 were classified as unengaged, 8 as proactive, and 23 as reactive. Prior to conducting the test, we assessed the reliability of the employed scales by calculating their corresponding Cronbach's Alpha coefficient. In all the cases, such a value was higher than 0.7 suggesting a high internal consistency within each scale's items (Gliem and Gliem 2003).

Table 5 also summarises the outcome of the one-way ANOVA for each constructs measured. We found no significant differences in any of these constructs across proactive, reactive, and unengaged users. This was also the case when conducting a *t*-Test for a two-group classification (i.e. engaged and unengaged).

5 Discussion

This section discusses the results of our study and provides answers to the paper's research questions. We also elaborate on the implications of our findings within the area of developer-centred security, namely the elaboration of strategies for boosting the participation of SO users in cybersecurity discussions.

Table 5 One-way ANOVA Test (profile and survey data)

Variable	SS	d.f.	MS	F	*p*	η^2
Profile data						
% self-disclosure	9.340	2	4.670	86.180	0.000	0.024
Survey data						
GPC	0.825	2	0.412	0.333	0.718	0.011
PCST	0.400	2	0.200	0.141	0.869	0.005
PCOT	3.860	2	1.930	1.127	0.331	0.036
RSK	0.547	2	0.274	0.384	0.682	0.012
PC	2.174	2	1.087	0.854	0.431	0.027
SD	3.082	2	1.541	1.040	0.360	0.033

5.1 Engagement and self-disclosure behaviour (RQ1)

Our findings suggest that SO users with a tendency towards starting cybersecurity discussions disclose significantly less information in their profiles than others who do not (Sect. 4.2). Similar observations were made by Kayes et al. (2015) in a study about peoples' engagement in the Q&A platform Yahoo! Answers. The authors found correlations between users' self-disclosure behaviour (i.e. profile visibility preferences), the frequency, and the quality of their contributions. Particularly, individuals with a more restrictive profile tend to contribute more and with better content than those with a public one. Furthermore, such users also showcase higher retention levels (i.e. average time interval between contributions) and have a higher perception on answer quality.

On the other hand, our results also show that *reactive* users not only reveal more profile information than *proactive* ones, but also more than those *unengaged*. Such a finding is to some extent aligned with prior research on identity formation in Q&A platforms. To a certain extent, participation in SO is driven by users' need for recognition within the platform. That is, in terms of points and badges that users can assign to each other based on the perceived quality of their contributions (Yang et al. 2014). For instance, a study conducted by Adaji and Vassileva (2016) showed that high-quality questions are frequently posted by users with complete profile information. Vargo and Matsubara (2018) also made similar observations and concluded that profile visibility tends to decrease over time. Hence, we could assume that reactive users may also be driven by reputation or recognition when deciding whether to disclose more personal information inside their profiles.

5.2 Engagement and privacy-related constructs (RQ2)

Unlike the results obtained from the users' profile information (Sect. 4.2), the analysis conducted over the survey data showed no significant differences in the elicited constructs (i.e. GPC, PCST, PCOT, RSK, and PC) across *unengaged*, *proactive*, and *reactive* users (Sect. 4.3). We hypothesise that this can be related to the relatively good reputation of SO in terms of privacy and data protection, as opposed to OSNs like Facebook. Unlike the latter, SO has not received the attention of mainstream media due to major data-breach scandals or privacy violations. Hence, the role of privacy concerns and perceived risks may not be significant for users' participation and engagement within the platform.

The differences observed in self-disclosure behaviour were not reflected by its survey counterpart (i.e. the SD variable). Nevertheless, and despite that such results may look inconsistent, prior research has also found discrepancies between people's *reported* and *actual* privacy behaviour. As mentioned in Sect. 2.2, this is often referred to as the "privacy paradox", a phenomenon frequently observed within users of OSNs. Our findings suggest traces of this paradox among SO users, especially when contrasting the outcome of the survey analysis with that of the users' profiles. Still, further research is necessary to determine whether the reported privacy behaviour outperforms the actual one across the three user categories. It would be of special interest to understand whether and up to which extent is the privacy paradox manifested among SO users, and how does it relate to their overall engagement.

5.3 Implications and recommendations

As privacy and security flaws in information systems grow steadily, it is very important to promote the exchange of privacy and security knowledge among software practitioners. To a large extent, the SO community is encompassed by early-career developers seeking for support and guidance in their engineering practices (Lopez et al. 2018). Hence, it plays a key role in the dissemination and synthesis of cybersecurity expertise. However, our results show an apparent deficit in terms of answers to privacy and security-related questions (Sect. 4.1). This can not only cause dissatisfaction to those asking such questions, but also damage the platform's value and usefulness in this regard.

Having identified nuances in the self-disclosure behaviour across different user groups can be used to foster the exchange of privacy and security expertise. For instance, profile information could be leveraged to motivate the participation in cybersecurity discussions among SO users, by rooting pending questions to those users who are more likely to answer them (e.g. those with a less visible profile). Moreover, closed questions could be assigned to these users for further clarification, and thus increase their resolution chances. Such an approach could also contribute to existing Q&A recommender systems and frameworks (e.g. Wang et al. 2016) seeking to match forthcoming questions to potential respondents. That is, by incorporating profile visibility as a feature of their question-user matching algorithms.

Similarly, our results could be used to elaborate incentive strategies targeting unengaged individuals. For example, by delivering cybersecurity suggestions to those SO users having a more visible profile. This approach is illustrated in Fig. 4, where a (hypothetically) unengaged user

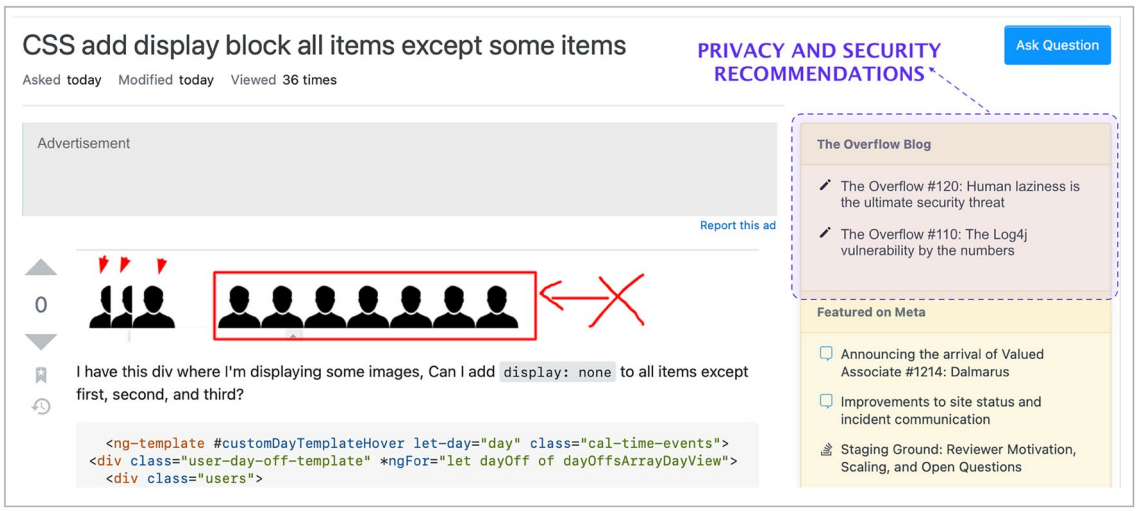

Fig. 4 Practical implications (envisaged interface)

receives such suggestions as she seeks for advice about an issue that is not cybersecurity-related. Here suggestions come in the form of privacy and security-related entries in the Overflow Blog,[12] a website curated by SO that gathers essays, opinion articles, and podcasts about computer programming. Using different persuasive styles to approach certain user groups could also improve even more the chances of engagement and behaviour change (Aimeur et al. 2019; Schäwel 2019). For example, unengaged users could be nudged using a more *authoritarian* style (e.g. *"Microsoft and other big tech companies urge developers to engage in cybersecurity training!"*), whereas a *consensual* one could be applied to proactive and reactive users (e.g. *"Many across the SO community agree: Cybersecurity training is essential for software developers!"*). Likewise, differentiated training content (e.g. access to customised documentation and software artifacts) could be offered to each user group based on a further assessment of their technical skills.

6 Limitations

To a certain extent, the results of our study are subject to limitations related to its experimental design. One is related to the methodology employed to compute users' amount of self-disclosure. Particularly, it should be noted that profiles are not the only means to reveal private information in SO as users can also disclose personal data inside questions, answers, or comments. However, we conducted our analysis exclusively over SO profiles as they are already adequate and extensive sources of self-disclosure evidence. Likewise, the approach we followed to characterise users' engagement is subjected to limitations. Indeed, engagement can also take a passive form, where a member (often referred as a "lurker" Oliveira et al. 2018) may not contribute actively to a discussion but may still read it and take advantage of its knowledge. We left passive engagement out of the scope of this work as it cannot be determined from the information in our dataset. Still, future research will seek to characterise lurkers and their interaction patterns regarding cybersecurity discussions.

Some characteristics of the studied sample (i.e. discussions and profiles) may also affect the generalisability of the discussed results. Particularly, our selection of cybersecurity questions was guided exclusively by the tags users assign to them. Hence, we may have considered wrongly-tagged questions in our analysis or missed some untagged ones out. Nonetheless, recent research in tag recommendation suggest that the number of wrongly-tagged content in SO remains relatively small (He et al. 2022). Moreover, since the SO community of curators often addresses such inconsistencies,[13] we assumed the posts we retrieved were accurately labelled. In the same regard, the different sample sizes between Python and R discussions can be considered a threat to the external validity of our results. We have addressed this issue by treating both samples as one without conducting any analysis on each specific language.

Another shortcoming stems from the approach we followed for distributing the survey. Overall, such an approach can lead to a "survivorship" bias as the survey was only distributed among those SO users whose e-mail addresses were

[12] https://stackoverflow.blog

[13] https://stackoverflow.com/help/privileges/suggest-tag-synonyms

retrieved successfully from their profiles (i.e. from both SO and GitHub). Hence, this may provide insights from users prone to disclose their e-mail addresses in these platforms but not on those involved in cybersecurity questions. Moreover, as shown in Table 1, survey respondents were predominately men which, despite reflecting current demographic trends in SO (Ford et al. 2017), offers a narrow view over the analysed behaviour. Convenience sampling methods like this one are pretty popular when conducting empirical studies in software engineering (Baltes and Diehl 2016). This is mainly due to the hardships of gathering empirical insights beyond profile information and discussion threads. However, these sampling methods often fail to draw a complete picture of the investigated phenomena (Baltes and Ralph 2022). We sought to mitigate the counter-effects of this approach by reaching out to as many SO members as possible. Nevertheless, larger and more gender-diverse samples would be necessary for the sake of generalisability.

Finally, having analysed the connection between users' self-disclosure practices and their engagement in cybersecurity discussions offers just a partial view on a multifaceted phenomenon. As mentioned in Sect. 5.1, both self-disclosure and engagement practices can be influenced by users' need for recognition and popularity within the platform, among other intrinsic and extrinsic factors. Hence, we acknowledge that our study is observational and, as such, cannot be leveraged to draw casual conclusions given the lack of controlled experimental ground truth data.

7 Conclusions and future work

Secure software development largely depends on practitioners' abilities to detect and address potential cybersecurity threats. Still, prior work has shown that many consider security and privacy as secondary aspects of software projects (Acar et al. 2016). Given the increasing popularity of Q&A platforms like SO, it is important to characterise and foster the exchange of cybersecurity expertise of their users in order to shape privacy- and security-savvy communities.

The results of this work confirm that "*answer-hungry*" questions are still a pending issue in SO. Furthermore, it is an issue affecting the privacy and security-related expertise provided by the platform and its community. As discussed in Sect. 5.3, having identified different engagement patterns can contribute to elaborating recommender systems and incentive mechanisms targeting this issue. Considering SO's size and outreach, these results could also support the dissemination of privacy- and security-by-design principles among software practitioners. That is, by delivering personalised training programs and tools through the platform to bridge developers' knowledge gaps on cybersecurity. Hence, this work contributes not only to current research in SO but also to ongoing efforts on bringing cybersecurity to the core of software engineering practices.

As highlighted in Sect. 6, the results yielded in this work are observational and call for further investigations. One potential direction for future research is the interplay between privacy concerns and the quality of cybersecurity feedback provided by SO users. For instance, to determine whether developers' *collective privacy concerns* (e.g. their sense of responsibility and empathy towards end-users) and prior cybersecurity experience play a significant role in the extent and frequency of their contributions. For this, we plan to extend our analysis with an empirical study about the factors motivating developers to value and address security and data-protection aspects of their software. For instance, by using scales and survey instruments that capture their efforts towards secure software development, experiences with security issues along with extrinsic motivations and deterrents (similar to the ones proposed in (Assal and Chiasson 2019) and (Tahaei et al. 2021)).

Appendix A: card sorting

Display names in SO are kept separately from the real private name. While the former is publicly shown in the network, the latter is used only on SO-directed job applications and remains inaccessible through the StackExchange API. *Locations* are also presented as text fields allowing users to obfuscate this information (e.g. put their country instead of the city) or even write whatever they want. We conducted a card-sorting analysis to determine the reliability of these two fields (and of the proposed coding schema).

Card sorting is a frequently used technique to derive taxonomies and prevent biases when categorising data (Whitworth et al. 2006). During a card sorting iteration, a person organises entities into a set of prefixed categories based on some common criterion. Once two or more people conduct this process, classification disagreements are discussed and resolved consensually. Here, each author performed a single card sorting iteration (i.e. one for display names and one for locations) using the following classification rules:

- *Display names* were categorised as **1** in case of full real names, **0.5** for partial or shortened real names, and **0** in the case of fantasy names.
- *Locations* were classified as **1** when including a ZIP/Postal code, **0.66** for cities or states, **0.33** for countries or large regions (e.g. Europe), and **0** in case of fictitious places.

We conducted such an analysis over a generalisable sample of locations and display names using a confidence level of 95%. Since *display names* are mandatory, we generated a

sample of 374 names from a population of 3574. On the other hand, we sampled 329 locations out of a total of 2248 (we only considered those that were not empty). Both samples were randomly selected using an R script.

Once each author had sorted the samples, we applied Cohen's Kappa coefficient to compute the inter-rater reliability of both classifications (McHugh 2012). Cohen's Kappa measures the level of agreement between two or more raters responsible for sorting items into mutually exclusive categories. A value closer to +1 suggests a high inter-rater agreement, whereas a value approaching −1 would indicate a high disagreement among raters. We adopted an agreement/disagreement threshold of 0.79 according to common practices and conventions for its adequate interpretation (McHugh 2012).

The coefficients obtained in both cases (+0.80 for *display names* and +0.86 for *locations*) indicate a high rate of agreement and reliability of the corresponding coding schemas. Moreover, after completing the card sorting analysis, 55% of the *display names* had been classified as full or partially-full real names, and 97% of locations as cities, states, countries, or large regions. **Hence, we concluded that when such information is present in a SO profile, it can be considered accurate.**

Author contributions NEDF: Conceptualization, Methodology, Investigation, Writing- Original Draft Preparation, Formal Analysis, Data Curation. MV: Conceptualization, Methodology, Investigation, Validation, Writing- Original Draft Preparation, Data Curation. MH: Validation, Writing- Reviewing and Editing. RS: Validation, Writing-Reviewing and Editing.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare no competing interests.

References

Acar Y, Fahl S, Mazurek ML (2016) You are not your developer, either: A research agenda for usable security and privacy research beyond end users. In: Cybersecurity Development, pp 3–8. IEEE, Boston, MA, USA

Adaji I, Vassileva J (2016) Towards Understanding User Participation in Stack Overflow Using Profile Data. In: International Conference on Social Informatics, pp 3–13. Springer, USA

Ahmad A, Feng C, Ge S, Yousif A (2018) A survey on mining stack overflow: question and answering (Q&A) community. Data Technol Appl 52(2):190–247

Ahmed T, Srivastava A (2017) Understanding and evaluating the behavior of technical users: a study of developer interaction at StackOverflow. Human-Centric Comput Inf Sci 7(1):1–18

Aïmeur E, Diaz Ferreyra NE, Hage H (2019) Manipulation and malicious personalization: Exploring the self-disclosure biases exploited by deceptive attackers on social media. Front Artif Intell 2:26

Assal H, Chiasson S (2018) Motivations and amotivations for software security. SOUPS Workshop on Security Information Workers (WSIW). USENIX Association. USENIX Association, USA, pp 1–12

Assal H, Chiasson S (2019) 'Think secure from the beginning' A Survey with Software Developers. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp 1–13

Baltes S, Diehl S (2016) Worse than spam: Issues in sampling software developers. In: Proceedings of the 10th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, pp 1–6

Baltes S, Ralph P (2022) Sampling in software engineering research: a critical review and guidelines. Empir Softw Eng 27(4):1–31

Buchanan T, Paine C, Joinson AN, Reips U-D (2007) Development of measures of online privacy concern and protection for use on the Internet. J Am Soc Inf Sci Technol 58(2):157–165

Choi TR, Sung Y (2018) Instagram versus snapchat: self-expression and privacy concern on social media. Telemat inform 35(8):2289–2298

Chua AY, Banerjee S (2015) Answers or no answers: studying question answerability in stack overflow. J Inf Sci 41(5):720–731

European Comission (2021) Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). European Comission

Fischer F, Böttinger K, Xiao H, Stransky C, Acar Y, Backes M, Fahl S (2017) Stack Overflow Considered Harmful? The Impact of Copy &Paste on Android Application Security. In: Symposium on Security and Privacy (SP), pp 121–136. IEEE, USA

Ford D, Harkins A, Parnin C (2017) Someone like me: How does peer parity influence participation of women on Stack Overflow? In: 2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pp 239–243. IEEE

Gao Z, Xia X, Lo D, Grundy J (2020) Technical Q &A site answer recommendation via question boosting. ACM Trans Softw Eng Methodol (TOSEM) 30(1):1–34

Garson GD (2012) Testing Statistical Assumptions. Statistical Publishing Associates, Asheboro, NC 27205 USA

Garson GD (2014) Logistic Regression: Binary and Multinomial. Statistical Publishing Associates, Asheboro, NC 27205 USA

Gliem JA, Gliem RR (2003) Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. In: 2003 Midwest Research to Practice Conference in Adult, Continuing, and Community Education

Hadar I, Hasson T, Ayalon O, Toch E, Birnhack M, Sherman S, Balissa A (2018) Privacy by designers: software developers' privacy mindset. Empir Softw Eng 23(1):259–289

He J, Xu B, Yang Z, Han D, Yang C, Lo D (2022) Ptm4tag: sharpening tag recommendation of stack overflow posts with pre-trained models. In: Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension, pp 1–11

Jozani M, Ayaburi E, Ko M, Choo K-KR (2020) Privacy concerns and benefits of engagement with social media-enabled apps: a privacy calculus perspective. Comput Human Behav 107:106260

Kayes I, Kourtellis N, Bonchi F, Iamnitchi A (2015) Privacy Concerns vs. User Behavior in Community Question Answering. In: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp 681–688. IEEE, Boston, MA, USA. IEEE

Krämer NC, Schäwel J (2020) Mastering the challenge of balancing self-disclosure and privacy in social media. Curr Opin Psychol 31:67–71

Krasnova H, Günther O, Spiekermann S, Koroleva K (2009) Privacy concerns and identity in online social networks. Identity Inf Soc 2(1):39–63

Kung FYH, Kwok N, Brown DJ (2018) Are attention check questions a threat to scale validity? Appl Psychol 67(2):264–283

Lopez T, Tun T, Bandara A, Mark L, Nuseibeh B, Sharp H (2019) An Anatomy of Security Conversations in Stack Overflow. In: 41st International Conference on Software Engineering: Software Engineering in Society, pp 31–40. IEEE/ACM, Canada

Lopez T, Tun TT, Bandara A, Levine M, Nuseibeh B, Sharp H (2018) An Investigation of Security Conversations in Stack Overflow: Perceptions of Security and Community Involvement. In: 1st International Workshop on Security Awareness from Design to Deployment. SEAD '18. ACM, USA, pp 26–32

McHugh ML (2012) Interrater reliability: the Kappa statistic. Biochem Med 22(3):276–282

Moutidis I, Williams HT (2021) Community evolution on stack overflow. Plos one 16(6):0253010

Oliveira N, Muller M, Andrade N, Reinecke K (2018) The exchange in StackExchange: Divergences between Stack Overflow and its culturally diverse participants. In: Proceedings of the ACM on Human-Computer Interaction 2(CSCW), 1–22

Parliament E (2016) of the Council: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. Official Journal of the European Union (OJ) 59(1–88):294 (**European Comission**)

Ross A, Willson VL (2017) Independent Samples T-Test, pp 21–24. SensePublishers, Rotterdam. https://doi.org/10.1007/978-94-6351-086-8_5

Ross A, Willson VL (2017b) One-Way Anova, pp 21–24. SensePublishers, Rotterdam. https://doi.org/10.1007/978-94-6351-086-8_5

Schäwel J (Nov 2019) How to raise users' awareness of online privacy. PhD thesis, University of Duisburg-Essen

Seamons K (2022) Privacy-Enhancing Technologies. In: Modern Socio-Technical Perspectives on Privacy, pp 149–170. Springer, Cham

Senarath A, Arachchilage NAG (2018) Why Developers Cannot Embed Privacy into Software Systems? An Empirical Investigation. In: 22nd International Conference on Evaluation and Assessment in Software Engineering 2018, pp 211–216. ACM, USA

Sengupta S, Haythornthwaite C (2020) Learning with comments: An analysis of comments and community on Stack Overflow. In: Proceedings of the 53rd Hawaii International Conference on System Sciences

Sirur S, Nurse JRC, Webb H (2018) Are we there yet? understanding the challenges faced in complying with the general data protection regulation (gdpr). In: 2nd International Workshop on Multimedia Privacy and Security. MPS '18. ACM, USA, pp 88–95

StackExchange: Stack Overflow Statistics. https://stackexchange.com/sites

StackExchange: How Many Developers Visit Stack Overflow? https://stackoverflow.co/advertising/audience/

StackOverflow: Stack Overflow Tag Explorer. https://stackoverflow.com/tags

StackOverflow: Stack Overflow Tag Explorer. https://stackoverflow.com/help/privileges/suggest-tag-synonyms

StackOverflow: Stack Overflow - Where Developers Learn, Share, and Build Careers. https://stackoverflow.com

StackOverflow: The Overflow - Essays, Opinions, and Advice on the Act of Computer Programming from Stack Overflow. https://stackoverflow.blog

Staddon J, Huffaker D, Brown L, Sedley A (2012) Are Privacy Concerns a Turn-off? Engagement and Privacy in Social Networks. In: Proceedings of the Eighth Symposium on Usable Privacy and Security. SOUPS '12. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2335356.2335370

Tahaei M, Frik A, Vaniea K (2021) Privacy Champions in Software Teams: Understanding Their Motivations, Strategies, and Challenges. In: Conference on Human Factors in Computing Systems, pp 1–15. ACM, USA

Tahaei M, Vaniea K, Saphra N (2020) Understanding Privacy-Related Questions on Stack Overflow. In: Conference on Human Factors in Computing Systems, pp 1–14. ACM, USA

Tahaei M, Li T, Vaniea K (2022) Understanding privacy-related advice on stack overflow. Proc Priv Enhanc Technol 2022(2):114–131

Vargo AW, Matsubara S (2018) Identity and performance in technical Q &A. Behav Inf Technol 37(7):658–674

Wang L, Wu B, Yang J, Peng S (2016) Personalized recommendation for new questions in community question answering. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp 901–908. IEEE, Boston, MA, USA. IEEE

Whitworth B, Ahmad A, Soegaard M, Dam R (2006) Encyclopedia of Human Computer Interaction. von C. Ghaoui. Hershey: Idea Group Reference. Kap. Socio-technical systems 1(1):533–541

Yang J, Tao K, Bozzon A, Houben G-J (2014) Sparrows and Owls: Characterisation of Expert Behaviour in Stack Overflow. In: International Conference on User Modeling. Adaptation, and Personalization. Springer, Denmark, pp 266–277