Towards Privacy and Security in Private Clouds: A Representative Survey on the Prevalence of Private Hosting and Administrator Characteristics

Lea Gröber, *CISPA Helmholtz Center for Information Security and Saarland University;*
Simon Lenau and Rebecca Weil, *CISPA Helmholtz Center for Information Security;*
Elena Groben, *Saarland University;* Michael Schilling and Katharina Krombholz,
CISPA Helmholtz Center for Information Security

https://www.usenix.org/conference/usenixsecurity24/presentation/gröber-private-clouds

This paper is included in the Proceedings of the
33rd USENIX Security Symposium.

August 14–16, 2024 • Philadelphia, PA, USA

978-1-939133-44-1

Open access to the Proceedings of the
33rd USENIX Security Symposium
is sponsored by USENIX.

Towards Privacy and Security in Private Clouds: A Representative Survey on the Prevalence of Private Hosting and Administrator Characteristics

Lea Gröber [†‡] Simon Lenau[†] Rebecca Weil[†] Elena Groben[‡]
Michael Schilling[†] Katharina Krombholz[†]

[†]*CISPA Helmholtz Center for Information Security*
[‡]*Saarland University*

Abstract

Instead of relying on Software-as-a-Service solutions, some people self-host services from within their homes. In doing so they enhance their privacy but also assume responsibility for the security of their operations. However, little is currently known about how widespread private self-hosting is, which use cases are prominent, and what characteristics set self-hosters apart from the general population. In this work, we present two large-scale surveys: (1) We estimate the prevalence of private self-hosting in the U.S. across five use cases (communication, file storage, synchronized password managing, websites, and smart home) based on a representative survey on Prolific ($\mathbf{n = 1505}$). (2) We run a follow-up survey on Prolific ($\mathbf{n = 589}$) to contrast individual characteristics of identified self-hosters to people of the same demographics who do not show the behavior.

We estimate an upper bound of 8.4% private self-hosters in the U.S. population. Websites are the most common use case for self-hosting, predominately running on home servers. All other use cases were equally frequent. Although past research identified privacy as a leading motivation for private self-hosting, we find that self-hosters are not more privacy-sensitive than the general population. Instead, we find that IT administration skills, IT background, affinity for technology interaction, and "maker" self-identity positively correlate with self-hosting behavior.

1 Introduction

In the past decade, both end users and companies have migrated to public clouds [23, 28, 44, 61, 69]. Due to an abundance of Software-as-a-Service (SaaS) offerings, clouds are not only used for file storage but also for a broad set of use cases, effectively putting users' data in the hands of third parties. As an alternative, some people set up and maintain their own services on hardware they control. This behavior has distinct security and privacy implications for these so-called "self-hosters": (1) There is a privacy benefit in taking control of data by hosting services on controlled premises [32, 62, 63]. Nonetheless, this benefit is negated if the self-hosted service lacks proper configuration, maintenance, and backup procedures, thereby rendering it vulnerable to attacks or data loss. Accordingly, (2) self-hosters need to take responsibility for securing their operations, a task that requires technical knowledge that not all people who self-host have [26]. Some see a security advantage in self-hosting because they believe small instances are unattractive targets [26], while cloud providers are more prone to be attacked due to the centralization of user data [45]. Yet, big providers usually have the means to invest in experts to secure their operations. Having the means still does not imply that commercial clouds are secure [31, 48, 70] or that individuals without these means fail to secure their self-hosted services [46].

Currently, assessing the security and privacy implications of self-hosting is challenging given the diverse interplay of configurations, use cases, and operator capabilities that have not yet been thoroughly studied. Additionally, there is a lack of comprehensive data on the scale of self-hosting and how many people are impacted by its security and privacy implications. Moreover, to design context-specific solutions that make securing private data easy, we first need to understand the distinct disposition of the self-hoster population.

Accordingly, we propose the following research questions as first steps towards gauging the security impact of self-hosting and laying the foundations for designing effective solutions:

RQ1: *How widespread is self-hosting for private use cases? What kind of people are self-hosting?* Estimating the prevalence of self-hosting is a first step towards understanding its security and privacy impact. Characterizing self-hosters' demographics lays the foundations for constructing personas that are representative of the population and valuable for any kind of security-focused design efforts.

RQ2: *Which tools are self-hosted and how?* Understanding technologies and server-type choices enables us to concentrate research efforts on common security-

relevant and potentially high-risk use cases.

RQ3: *In which characteristics do self-hosters differ from the average U.S. population?* Investigating what distinguishes self-hosters from the general U.S. population contributes to a better understanding of enabling and constraining individual characteristics relevant to system administration work. It also hints at roadblocks in the hosting ecosystem.

We addressed these questions by using a sequential two-survey design. First, we identified self-hosters in a demographically representative U.S. sample with $n = 1505$ participants and analyzed which services they self-hosted. Second, we ran a follow-up survey with $n = 589$ participants and compared the identified self-hosters against a socio-demographically matched control group with regard to 14 characteristics suggesting a relationship with self-hosting. We selected characteristics that we hypothesized would predict self-hosting behavior, based on qualitative insights from focus groups we ran prior to survey construction. This work makes the following core contributions:

1. We estimate the prevalence of self-hosting for private use in the U.S. with an upper bound of 8.4 % CI [7, 9.6]. Based on this we suggest that self-hosting is not a niche phenomenon and that research efforts are worth investing to understand and support a population that is assumed to be strongly motivated by security and privacy concerns.

2. We compile demographic data and individual characteristics that describe the self-hoster population. This information may inform future design studies focused on developing security-enhancing solutions.

3. We provide an overview of prominent use cases and server-type choices highlighting potential high-risk scenarios. Understanding common self-hosting practices helps to identify critical use cases and informs future research and development efforts.

4. We highlight individual characteristics that are positively and negatively associated with self-hosting behavior. This yields insights about which factors could be roadblocks to self-hosting and helps concentrate efforts to support future administrators.

Replication Package. We provide a full replication package including survey instruments, anonymized data, and evaluation artifacts.[1]

2 Related Work and Self-Hosting Definition

Self-Hosting. Self-hosting falls within a spectrum of hardware as well as software control and maintenance responsibility. For the scope of this paper, we define self-hosting as (1) user control over the hardware, i.e., using own hardware or renting said hardware, (2) control over the software, i.e., the operating system, configuration, and applications, (3) the self-hosted service needs to be available over a network, and (4) responsibility for the service, e.g., not relying on a third party for set-up and maintenance.

The research community has been looking at self-hosting from the angle of people administrating their own home networks [6, 8, 20, 25, 65]. This line of work focuses on the growing complexity of home networks due to the increasing number of IoT devices. Problems arise when people have to deal with hardware and software failures and when expectations of usefulness are defied (e.g., mismatches between what a person expects to be able to do and specific device capabilities) [6]. Moreover, recent work investigated private and organizational self-hosting in the case of a popular file-sharing and content collaboration platform. Here, Gröber et al. [26] found that privacy, autonomy, and security are prominent motivational factors when people decide to self-host services for private use. These self-hosters are diverse in terms of technical background and face challenges when it comes to maintenance and security choices.

Security Practices of Administrators. We place our work in the broader context of the security and privacy challenges administrators face. Self-hosters become admins without necessarily having the relevant expertise, however, even experts face challenges when working with security-relevant technology. In this regard, various studies examined root causes and usability challenges for transport layer security misconfigurations [22, 36, 37]. Similarly, Kraemer et al. [35] shed light on the influence of human error on network administration and information security. Li et al. [40] uncover challenges administrators face when updating servers. Dietrich et al. [15] investigated system operators' perspectives on security misconfigurations. They find that mitigations often exist but are not put to effective use. Moreover, recent research started investigating the social embeddedness of system administration work. Kaur et al. [34] uncovered structural challenges marginalized genders face when working in IT administration. Gröber et al. [26] categorized five different constellations in which operators work together to maintain IT infrastructure across organizational and private use. They found that individual characteristics such as operators' level of expertise and use case requirements influence how system administration work is carried out. These works highlight the relevance of the human-centric approach and justify our focus on individual characteristics.

Representative Studies in Security. We join the ranks of representative user studies in the area of security and privacy enabling researchers to draw generalizable conclusions. Redmiles et al. conducted a series of representative studies, such as a U.S.-census-representative survey via Survey Sampling International panel to investigate how demographics, knowledge, and beliefs correlate with security advice and be-

[1] https://projects.cispa.saarland/lea.groeber/usenix24-sh-prevalence

havior of end users [51]. Further, Redmiles et al. ran a census-representative telephone survey to understand how socioeconomic status correlates with security incidents [52]. Finally, Redmiles et al. investigated if data gathered on Amazon Mechanical Turk (MTurk) about security and privacy issues generalize to a broader population [53]. To this end, the authors compared an MTurk sample with a census-representative web panel and a probabilistic telephone sample. They found that MTurk responses regarding security and privacy tasks are actually more representative of the U.S. population than the web panel. This study has been replicated by Tang et al. on MTurk and Prolific and was compared against a probabilistic survey obtained through a service provider [64]. All studies above focus on the U.S. population. Recent research broadened the scope to German citizens [56], and a large-scale international study compared representative samples from 12 countries on four continents [30]. Both studies obtained their sample through online panelists carried out by a service provider. Research showed that online sampling platforms (including the one we used) are suitable tools for representative studies.

We offer a reflection on our study design and share takeaways from preparing and running large-scale representative studies on technical topics in Appendix A.

3 Methods Overview

This work comprises two quantitative surveys that we conducted consecutively. Thereby the self-hoster classification of Survey 1 informed the sampling of Survey 2. This way we were able to recruit a demographically matched control group for the self-hosters and study which characteristics distinguish them from non-self-hosters. Our team of researchers constructed both survey instruments simultaneously in an iterative process starting with paper prototypes. We continuously reviewed them for survey flow, potential biases, survey length, and attention checks and refined survey items (e.g., wording) and measurement details (e.g., scale anchors) to minimize misunderstandings or complications for participants. Two in-depth cognitive walkthroughs where we critically reviewed paper prototypes and the final instruments served as an internal quality check. We outline the final wordings and survey flows in the replication package. Both surveys also underwent technical pre-testing to ensure data collection quality and a smooth experience for participants.

We preregistered the study design including hypotheses.[2] Hypotheses for Survey 2 were grounded in qualitative data of two focus groups we ran prior to survey construction. Thus, we explain the focus groups as a methodological aspect of Survey 2 in Section 7.1.1. In the following, we present both studies in sequential order including details on methodology, results, and discussion.

[2] https://osf.io/4apwe/?view_only=b08a9b2d7b6d4f288b57f8382b26e41f

4 Prevalence Survey 1 - Methods

To allow for valid estimation of the prevalence of self-hosting behavior in the U.S., we aimed to recruit a representative sample of the U.S. population and administer a survey based on which responses we could identify self-hosters. Representativeness in the context of our study refers to meeting the census distribution of age, sex, and ethnicity and, due to sampling via Prolific, is restricted to individuals who have internet access via a computer or mobile device. We assume the influence of this restriction to be minimal because access to such devices exceeded rates of 94% for heads of household up to 64 years old in 2018. For heads of household aged 65 and older, the rate was 80% [43]. With a continuous positive trend in recent years, the influence today can be assumed to be even less [57]. There might be other restrictions (beyond device and internet access) with respect to representativeness we can only speculate about. For example, data collected with Prolific suffer less from unattentive, low-motivated participants compared to other online platforms [17], but we do not have data about distributions of, e.g., motivations or attentiveness of the population and can thus not adjust for it. Each form of assessment (e.g., online, face-to-face, telephone) may introduce biases [55], and representativeness is limited to the assessment methods, their confounds and the selected characteristics for representativeness. Accordingly, our results should be interpreted as such. Our measurements, sampling method, and identification strategy are detailed in the following.

4.1 Measurements

We introduced our survey as a part of research about software and application use in private and professional contexts to avoid biasing participants toward the subject of self-hosting. As a basis for an operational identification strategy of self-hosting behavior, we presented participants with 5 use case categories (i.e., file storage - synchronization, file transfer; Web sites - CMS, blogging; communication - messaging, voice/video telephony; synchronized password managing; smart home) and lists of 5 self-hostable and 6 non-self-hostable tools within each category (see Table 3), in random order. We determined use cases with a crowd-sourced list "Awesome-Selfhosted" on GitHub and narrowed them down according to focus group insights and our research questions. "Smart-home" is included as an edge-case for devices that are available over the network and connected to a self-hosted home automation service in a server-client model. We included tools based on their popularity on GitHub for self-hostable tools [7, 27] and on Google Trends (https://trends.google.com) for non-self-hostable tools. We defined exclusion criteria to ensure self-hostable software tools match our definition of self-hosting. The replication package contains a detailed write-up of the process. In the survey,

participants also had the opportunity to add other tools not mentioned in the pre-selected lists. We asked them to indicate for each tool whether they use it in a private or work context.

For each self-hostable tool, and for each tool entered via the "other" option that was being used in a private context, we asked participants whether the tool was set up for them or whether they had set it up themselves on a server. All participants indicating that they had set up the tool themselves were subsequently asked on which type of server they had set it up (i.e., home server, virtual private server, dedicated server, or other). This was identical for all use case categories, except the smart home category. Here, we asked participants if they had enabled remote access from outside the local network (i.e., accessible via the internet, accessible via a virtual private network, only on the local network).

To reduce the number of false positives (i.e., participants incorrectly identified as self-hosters based on their response in the tool selection) we introduced a definition of private self-hosting in the second part of Survey 1. We distinguished it from commercial cloud services and asked participants if they had come into contact with self-hosting before, providing us with a short description of how they had come into contact via a free text field. We used this information later on as a sanity check for the identification of self-hosters based on their tool selection. To further ensure data quality, we used two attention checks in the survey to identify inattentive participants [47].

In the last part of Survey 1, we collected demographic data and final comments on the survey, and we announced the follow-up survey (Survey 2) for which we might contact some of the participants again. Compensation was estimated based on Prolific's best practice (>= 9 GBP/hour) and pretest duration (~360 seconds). The median response time of 499 seconds ended up higher than our estimate, resulting in a lower hourly wage of 6.49 GBP.

4.2 Sampling

Based on the data available to us at the time of sampling in the U.S. Decennial Census of Population and Housing [67] we defined target sub-samples, representing the U.S. population in terms of sex, age, and ethnicity. Because census data only contain information about the distribution of sexes (i.e., male, female) but not gender, we sampled based on the information available to us. However, because we assessed self-reported gender (i.e., man, including trans man/trans male; woman, including trans woman/trans female; non-binary; self-described) in our surveys, we report this information in the result sections. Moreover, we used the brackets 18-28; 28-38; 38-48; 48-58, and $>$ 58 for the target age and Asian, Black, Mixed, Other, and White for the target ethnicity.

For data collection we used Prolific (www.prolific.co), defining 50 target sub-samples for U.S. residents as a result of the cross combination with sex, age, and ethnicity, utilizing the Prolific pre-screening functionality. Replicating Prolific's in-house representative sampling solution determined the simplification of ethnicity groups in the U.S. census [64, 66]. Prolific allows a relatively cost-effective and fast way to collect data that has been shown to be representative of the U.S. population, at least with respect to some assessed items [64]. As such it might be considered a more feasible approach for non-commercial research purposes as compared to other face-to-face omnibus representative sampling procedures (e.g., ipsos.com). Nevertheless, we have to assume a general bias in our sample, potentially relevant to our research questions, in the sense that it only contains people who have internet access via a computer or mobile device to answer Prolific surveys. Accordingly, we treat our estimation of the prevalence of self-hosting behavior in the U.S. population as an upper bound. The total targeted sample size was $n = 1500$, based on the availability in the Prolific participants pool. Data collection took place between August and September 2022.

4.3 Data Cleaning and Preparation

We only included participants in the sample who had completed the entire survey and passed both attention checks. Due to our aim to achieve a representative sample, we excluded participants who did not meet these criteria, and we re-sampled in line with the requirements of our target sub-samples. Due to technical issues, some participants were observed twice. For these participants, we kept only the first complete data set. Once data collection was complete, three researchers coded participants' final survey comments (codes: "issue", "no issue") to see whether any comments raised doubts about data quality, e.g., the participant mentioned having accidentally indicated to self-host a tool. In case of disagreement, they discussed their ratings and came to a consensus. Four participants were excluded from the sample.

4.4 Survey Weighting

The sampling procedure closely aligned the distribution of sex, age, and ethnicity in our sample to the distribution in the population (see Table 7). Still, slight deviations occurred due to integer sample size limitations and data cleaning (see section 4.3). To resolve these small discrepancies and accurately estimate self-hosting prevalence, we applied the Generalized Regression (GREG) Estimator [14] to obtain calibration weights for our survey, as is best practice in other research areas and official statistics [11,39,50,54]. The weights were determined such that the estimated weighted proportions exactly meet the census proportions in Table 7. The weights are $w_h = N_h \Big/ n_h^{(1)}$ for each element in the h-th socio-demographic group defined by sex, age, and ethnicity. N_h and $n_h^{(1)}$ denote the group sizes in the census and our first survey, respectively. This ensures that our results accurately reflect the diversity of socio-demographic groups in the population, mirroring their exact

proportions in the U.S. population. For example, this counterbalances variations among different groups in participant exclusions based on attention checks.

4.5 Operational Classification of Self-Hosters

We identified participants as self-hosters when they selected or added at least one self-hostable tool in one of the use case categories as privately used, indicated that they had set it up themselves on a server, and confirmed that they had come into contact with self-hosting before taking part in our survey. Alternatively, they were identified as self-hosters when they described in the open response format (i.e., Are there any other tools that you self-host and that have not been mentioned above?) that they use self-hosted tools or network services, set up and maintain services themselves, and are in control of the infrastructure. In addition, they had to confirm that they had come into contact with self-hosting before taking part in our survey. As the latter strategy relied on open responses, four raters coded the open responses independently, checking for the criteria described above. In case of disagreement, they discussed their ratings and came to a consensus. A final sanity check was applied to all self-hoster classifications. Open responses of all identified self-hosters were coded by four coders for any indications raising doubts about the classification (e.g., a non-self-hostable tool was mentioned by the participant as being self-hosted). In case of disagreement, they discussed their ratings and came to a consensus. The classification process produced three outcomes. Participants were either classified as self-hosters, as non-self hosters, or as having an unclear self-hosting status (i.e., due to conflicting information about their self-hosting behavior). This procedure does not necessitate inter-rater-reliability calculation, as conflicts are resolved.

5 Prevalence Survey 1 - Results

We achieved a representative sample ($n = 1505$) of the United States, with respect to age, sex, and ethnicity, in line with the U.S. Decennial Census of Population and Housing Census [67] and the corresponding demographic data available on Prolific. A comparison of the respective shares of age, sex, and ethnicity for our survey and the population can be found in Table 7 in the Appendix.

We rely on confidence intervals (CIs) for assessing statistical significance because, unlike p-values, they quantify (im)precision of reported quantities. For statistically independent groups, non-overlapping CIs imply significance [58]. Where groups are not independent, we additionally used tests for dependent samples (Wald- for binary and t-Tests for continuous variables) to ensure the correctness of reported significances. CIs alone yielded identical conclusions to tests for dependent observations.

Table 1: Estimated self-hosting prevalence and group comparison by age, ethnicity, sex and gender (in %)

		Distribution in subgroups	
	Prevalence	Self-Hosters	Non-Self-Hosters
Age			
18 – 28	8.3 ± 3.2	18.4 ± 6.3	18.6 ± 0.6
28 – 38	10.9 ± 3.7	22.3 ± 6.6	16.5 ± 0.6
38 – 48	7.9 ± 3.2	17.1 ± 6.3	18.2 ± 0.6
48 – 58	11.6 ± 3.6	**26.1 ± 7.1**	**18.1 ± 0.7**
58	4.9 ± 2.1	**16.2 ± 6.2**	**28.6 ± 0.6**
Ethnicity			
asian	8.0 ± 6.0	4.8 ± 3.5	5.0 ± 0.3
black	10.0 ± 4.1	14.3 ± 5.5	11.7 ± 0.5
mixed	5.9 ± 7.1	1.5 ± 1.8	2.2 ± 0.2
other	7.4 ± 5.3	5.7 ± 3.9	6.5 ± 0.4
white	8.3 ± 1.6	73.8 ± 7.1	74.7 ± 0.6
Sex			
female	2.9 ± 1.2	**17.8 ± 6.6**	**55.1 ± 0.8**
male	14.3 ± 2.5	**82.2 ± 6.6**	**44.9 ± 0.8**
Gender			
woman	3.1 ± 1.2	**18.6 ± 6.7**	**53.6 ± 1.0**
man	14.1 ± 2.6	**79.9 ± 6.9**	**44.2 ± 1.0**
non-binary	0.0	**0.0**	**1.3 ± 0.6**
self-described	11.0 ± 20.3	0.8 ± 1.6	0.6 ± 0.4
not stated	18.5 ± 33.1	0.7 ± 1.5	0.3 ± 0.3

Boldened shares indicate significant differences between estimated self-hoster and non-self-hoster shares

± indicates the lower and upper bounds of the 95% confidence intervals

5.1 RQ 1: Prevalence of Private Self-Hosting

One of the main goals of Survey 1 was the determination of self-hosting prevalence among the U.S. population. In total, we identified $n = 124$ self-hosters in our sample, indicating an upper-bound prevalence of 8.4 %, CI [7, 9.6] of self-hosting behavior in the US. We also identified $n = 1355$ non-self-hosters and $n = 26$ received an unclear self-hoster status. Table 1 shows the self-hosting prevalence (i.e., estimated occurrence of self-hosters in the population) broken down by age, ethnicity, sex and gender in column 2. For example, we estimated a self-hosting prevalence of 11.6 % in the age group 48-58. We compared the share of age, ethnicity, sex and gender characteristics between the group of self-hosters (column 3) and non-self-hosters (column 4) to determine any significant differences between the two groups with regard to these characteristics (e.g., are self-hosters more likely than non-self-hosters to fall in age group 48-58?). Compared to non-self-hosters, the age group of 48-58 is significantly more frequent, while people older than 58 years are less frequent in the self-hosters sample. Moreover, men are more while women and non-binary people are less frequent among iden-

Table 2: Usage and hosting type shares of self-hostable tools (in % of all self-hosters)

Tool	Self-Hosted by		Server type							
			Home		VPS		Dedicated		Other / unknown	
Communication	20.2	± 9.1								
Teamspeak	12.0	± 7.3	64.5	± 22.7	11.6	± 15.1	23.9	± 20.4	0.0	
Mumble	8.0	± 6.2	63.3	± 28.6	18.1	± 22.7	18.5	± 23.2	0.0	
Rocket.Chat	2.7	± 3.6	32.0	± 52.2	68.0	± 52.2	0.0		0.0	
Jitsi Meet	2.6	± 3.6	65.8	± 54.0	0.0		34.2	± 54.0	0.0	
Mattermost	1.3	± 2.5	32.1	± 52.2	67.9	± 52.2	0.0		0.0	
Other	1.4	± 2.7	0.0		50.1	± 69.3	0.0		49.9	± 69.3
File Storage	17.7	± 8.8								
Nextcloud	4.1	± 4.6	74.9	± 43.5	0.0		25.1	± 43.5	0.0	
ownCloud	2.8	± 3.8	49.0	± 69.3	51.0	± 69.3	0.0		0.0	
SparkleShare	2.7	± 3.7	100.0		0.0		0.0		0.0	
Synthing	2.6	± 3.6	65.7	± 54.1	34.3	± 54.1	0.0		0.0	
Seafile	1.4	± 2.7	100.0		0.0		0.0		0.0	
Other	6.7	± 5.7	59.5	± 43.0	20.2	± 35.4	0.0		20.2	± 35.4
Synchronized PW Managing	16.3	± 8.4								
Vault-/bitwarden	6.8	± 5.7	73.3	± 22.4	13.7	± 17.6	13.0	± 16.9	0.0	
sysPass	2.7	± 3.7	33.0	± 53.1	67.0	± 53.1	0.0		0.0	
Passbolt	1.3	± 2.6	75.2	± 42.2	24.8	± 42.2	0.0		0.0	
Teampass	1.3	± 2.6	50.3	± 69.3	49.7	± 69.3	0.0		0.0	
Other	6.8	± 5.7	89.0	± 20.4	0.0		0.0		11.0	± 20.4
Websites	51.4	± 11.4								
WordPress	46.0	± 11.4	47.9	± 11.9	26.2	± 10.4	21.5	± 9.7	4.4	± 4.8
Ghost	2.7	± 3.7	49.9	± 49.0	50.1	± 49.0	0.0		0.0	
Cockpit	1.3	± 2.6	51.4	± 69.2	48.6	± 69.2	0.0		0.0	
Other	4.0	± 4.5	33.2	± 37.7	33.4	± 37.7	0.0		33.4	± 37.7
			Accessible from Internet							
			yes		via VPN		no		Other / unknown	
Smart Home	22.8	± 9.6								
Home Assistant	21.5	± 9.4	25.3	± 13.4	0.0		59.4	± 15.2		
Node RED	2.6	± 3.6	25.8	± 43.5	0.0		24.5	± 41.8		
WebThings Gateway	1.4	± 2.6	0.0		100.0		0.0			

± indicates the lower and upper bounds of the 95% confidence intervals

tified self-hosters. All other assessed demographic characteristics did not differ significantly between self-hosters and non-self-hosters.

5.2 RQ 2: Tool Usage Patterns

To better understand self-hosting behavior, we looked at the distribution of the different use cases and tools selected from the range of presented tools by participants identified as self-hosters. Tools additionally listed by these participants under a respective use case were, due to relatively small numbers, summarized as 'Other'. Participants who we identified as self-hosters solely based on their responses in other open response formats (i.e., Are there any other tools that you self-host and that have not been mentioned above?) are not included in the table due to inconsistent information about use cases, tool names, and server types. Table 2 shows that from our predefined use cases, websites are most frequent among self-hosters. Communication, file storage, synchronized password managing, and smart home are less frequent use cases and do not differ significantly from each other in their frequency. For websites, the most frequently used tool is WordPress, which in the majority of cases is hosted on a home server. For the smart home use case, Home Assistant is most frequently used and in the majority of cases not accessible from the

internet. All tools that were indicated as being privately used by self-hosters from our pre-selection of tools can be found in Table 2.

In addition, we not only looked at the tools that are self-hosted by participants but also at the usage of non-self-hosted tools (i.e., self-hostable and non-self-hostable) for the same use cases in self-hosters and non-self-hosters. Interestingly, Table 3 shows that, across all use cases, self-hosters seem to use more tools in general. That is, they not only use more self-hostbale tools but also more of the non-self-hostable tools (e.g., Microsoft Teams, Google Drive and Home, LastPass) as compared to non-self-hosters.

6 Prevalence Survey 1 - Discussion

The goal of Survey 1 was to determine the prevalence of self-hosting in the U.S. to get a better understanding of how widespread the phenomenon is. To this end, we used a representative sampling method (additionally corrected by weighting), which in turn allows us to gauge the upper bound of the occurrence of self-hosting in the U.S. population. Based on the results of Survey 1 we estimate the occurrence of self-hosting with an upper bound of 8.4 %, CI [7, 9.6]. As such, self-hosting should not be considered a niche phenomenon.

The results of Survey 1 also indicated that self-hosters are more likely to be male and in the 48-58 age group, and less likely to be older than 58. Speculatively, a possible connection between the age of the participants and self-hosting could be the time of the emergence of relevant technologies and the life phases in which the people were at that time. People in the 48-58 age group were born between 1964 and 1974. This means that they were in the late adolescent phase of their lives at the time of the advent of the Internet and might have been open to innovations. At that time, however, the Internet was more technically demanding. More technical knowledge was required, and cloud computing in the current sense did not exist back then. If you wanted a certain service or functionality, "self-hosting" was the default. This is one attempt to reason about this finding, but the present research approach does not allow us to verify such claims. Still, our finding serves as a basis for future research to explore causal demographic variables to explain self-hosting behavior.

Our results also revealed that a large proportion of self-hosting behavior is hosting WordPress websites on a home server. This is a potential high-risk use case, as people are running internet-connected services from their homes. Further research might look into this use case to investigate security configurations and assist people in making secure decisions. Notably, our analyses also showed that being a self-hoster does not necessarily mean solely turning to self-hostable tools and avoiding commercial alternatives. Quite the contrary, across all use cases, self-hosters seem to use more self-hostable and commercial tools in general as compared to non-self-hosters. An avenue for future research is to explore

Table 3: User share per pre-defined tool (in %)

	Tool	Usage				
		Self-Hosters		Non Self-Hosters		
Communication	Zoom	77.5	± 7.3	75.6	± 2.3	
	Discord	**67.8**	± **8.2**	**35.6**	± **2.3**	
	Microsoft Teams	**48.5**	± **8.8**	**38.8**	± **2.5**	
	Whatsapp	37.0	± 8.5	37.8	± 2.5	
	Telegram	**27.4**	± **7.9**	**15.7**	± **1.9**	
	Signal	**16.9**	± **6.6**	**7.2**	± **1.4**	
	Mumble	**10.4**	± **5.4**	**1.1**	± **0.6**	✘
	Teamspeak	**10.4**	± **5.4**	**2.6**	± **0.8**	✘
	Jitsi Meet	**5.7**	± **4.1**	**1.1**	± **0.6**	✘
	Rocket.Chat	4.9	± 3.8	1.2	± 0.6	✘
	Mattermost	3.2	± 3.1	0.8	± 0.5	✘
File Storage	Google Drive	**92.0**	± **4.7**	**78.6**	± **2.1**	
	Dropbox	**67.0**	± **8.3**	**53.0**	± **2.6**	
	Microsoft OneDrive	59.6	± 8.7	51.1	± 2.6	
	iCloud	46.7	± 8.8	49.0	± 2.6	
	MEGA	**24.2**	± **7.5**	**7.6**	± **1.3**	
	Box	12.2	± 5.8	7.1	± 1.4	
	Nextcloud	**8.2**	± **4.8**	**1.0**	± **0.5**	✘
	ownCloud	**5.8**	± **4.1**	**1.0**	± **0.5**	✘
	SparkleShare	**4.9**	± **3.8**	**0.9**	± **0.5**	✘
	Seafile	4.9	± 3.8	1.0	± 0.5	✘
	Syncthing	4.0	± 3.5	1.2	± 0.6	✘
Smart Home	Amazon Alexa	47.7	± 8.7	38.7	± 2.6	
	Google Home	**46.0**	± **8.8**	**25.1**	± **2.3**	
	SmartThings	**20.0**	± **6.9**	**6.9**	± **1.3**	
	Home Assistant	**13.7**	± **6.1**	**3.2**	± **0.9**	✘
	Apple HomeKit	7.2	± 4.6	4.2	± 1.1	
	Node RED	4.0	± 3.5	1.0	± 0.5	✘
	WebThings Gateway	2.5	± 2.8	0.9	± 0.5	✘
	Bosch Smart Home	2.4	± 2.7	1.2	± 0.6	
	Domoticz	1.7	± 2.3	0.8	± 0.5	✘
	Gladys	1.7	± 2.3	1.0	± 0.5	✘
	Vivint Home	1.7	± 2.3	1.8	± 0.7	
Synchronized PW Managing	iCloud Keychain	24.9	± 7.5	20.4	± 2.1	
	LastPass	**21.9**	± **7.3**	**9.1**	± **1.5**	
	1Password	**13.9**	± **6.1**	**3.6**	± **1.0**	
	Vault-/Bitwarden	**12.9**	± **5.9**	**3.5**	± **1.0**	✘
	Roboform	**9.8**	± **5.2**	**2.7**	± **0.9**	
	Keeper	5.7	± 4.1	2.2	± 0.8	
	Dashlane	4.9	± 3.8	2.3	± 0.8	
	Padloc	**4.8**	± **3.8**	**0.9**	± **0.5**	✘
	Passbolt	4.1	± 3.5	1.0	± 0.5	✘
	Teampass	4.1	± 3.5	1.6	± 0.7	✘
	sysPass	4.1	± 3.5	1.1	± 0.6	✘
Websites	WordPress	**48.4**	± **8.8**	**19.7**	± **2.1**	✘
	Blogger	**16.2**	± **6.5**	**8.1**	± **1.4**	
	Wix	10.5	± 5.4	7.0	± 1.4	
	Squarespace	8.9	± 5.0	7.6	± 1.4	
	Weebly	7.3	± 4.6	4.1	± 1.1	
	Ghost	**4.9**	± **3.8**	**0.8**	± **0.5**	✘
	Webflow	4.8	± 3.7	1.6	± 0.7	
	Strapi	2.5	± 2.7	0.7	± 0.5	✘
	Cockpit	2.4	± 2.7	0.6	± 0.4	✘
	Jimdo	2.4	± 2.7	0.7	± 0.4	
	Wagtail	2.4	± 2.7	0.7	± 0.4	✘

Boldened shares indicate significant differences between estimated self-hoster and non-self-hoster shares
✘: Tool is self-hostable
± indicates the lower and upper bounds of the 95% confidence intervals

usage dependencies between tools and across use cases when people self-host.

7 Characteristics Survey 2 - Methods

To be able to better describe the group of self-hosters and understand individual characteristics that correlate with self-hosting behavior, we invited all self-hosters identified with the help of Survey 1 to take part in Survey 2 and matched them with an invited control group of non-self-hosters identified in Survey 1.

7.1 Measurements

The median response time was 614.5 seconds, and participants were compensated with an hourly wage of 14.01 GBP. This section presents details of the survey instrument as well as the hypothesis construction based on qualitative insights from focus groups.

7.1.1 Focus Groups to Identify Predictors

To identify candidate characteristics relevant to self-hosting behavior we conducted two focus groups with self-hosters (three participants) and non-self-hosters (ten participants), respectively. We provide an overview of the procedure, analysis, and results below.

One researcher moderated both sessions, which took about 90 mins each. The researcher followed a protocol (detailed in the replication package) but allowed and encouraged participants to discuss topics freely. To provide a common ground for everyone, the sessions started with an introduction to self-hosting which was especially vital for the non-self-hoster group. As an ice-breaker question, we asked participants about their prior experiences with self-hosting. Only the self-hoster group was then asked about their personal definition of self-hosting, contrasting it with the definition we offered. Afterward, the researcher opened the main discussion which was structured into six key questions: We explored reasons that would discourage individuals from using cloud services (Q1) and why they might be inclined to engage in self-hosting (Q2). Then, participants reflected on situations that might have influenced their decision to self-host (Q3). Moreover, we discussed other possible aspects and domains of life that could be relevant to self-hosting, such as personality traits and individual characteristics (Q4). Last, participants reflected on possible reasons for not self-hosting even though one would want to (Q5) and why some individuals would reject to self-host altogether (Q6). Afterward, we asked only the self-hoster group to identify technical skills they consider essential for self-hosting.

Thematic Analysis. Two researchers (computer scientist and psychologist) first grouped the audio data by questions, then listened repeatedly while applying open coding [9,13,59] independent from each other. The researchers discussed their coding, resolving all mismatches by revisiting the audio data and updating the codes. During this iterative process, it became evident that coding across questions yielded a better fit with the data compared to strictly adhering to the question structures. For instance, the theme "privacy" was observed both as a lack of privacy, which acted as a deterrent from using cloud services, and as a desire to attain privacy, which served as a motivation for self-hosting. Consequently, the coders developed the codebook to capture such overarching concepts. Once the coders reached an agreement, they organized the resulting codes in a mindmap, grouping them into topics and illustrating connections between them. Finally, they used the mindmap as the basis for a discussion to identify themes. In doing so, the coders listened to audio data again, this time identifying and transcribing relevant quotes. They identified ten core themes, for which they found supporting data in both focus groups: *work effort* (the amount of work, time and resources it requires to self-host), *security* (security advantages and disadvantages of self-hosting), *technology interest and skills* (aptitude for and ability to acquire know-how to self-host), *tinkering and DIY* (aptitude for self-hosting and ability to acquire the respective know-how), *interpersonal aspects* (different personal factors influencing the motivation to self-host), *money* (financial aspects involved in self-hosting, required spendings and saving money), *privacy* (privacy concerns and needs that can be addressed by self-hosting), *usefulness of self-hosting* (fulfilling unique needs that are not fulfilled by other services), *control* (self-hosting as a means to gain control over one's own life), *openness to new things* (trying out things and setting trends). Many of these core themes concur with recent research findings, identifying privacy, security, and autonomy needs, saving costs, usefulness, and enjoying learning something new as motivational factors for self-hosting [26]. The authors also showed that a specific skill set and expertise is needed (or needs to be brought in) for self-hosting.

7.1.2 Scale Measurements and Hypotheses

We mapped scale measurements of individual characteristics to all core themes that could be captured by such measures and for which we found validated and reliable scales in the literature (i.e., *security, privacy, technology interest and skills, openness to new things, tinkering and DIY, money, work effort, control*). This allowed us to empirically test if the characteristics that emerged from the qualitative analysis of the focus groups can predict self-hosting behavior. In the survey, scale order was randomized. Similar to Survey 1, to ensure data quality, we used two attention checks in the survey to identify inattentive participants [47]. Below we provide details on the hypotheses and scales we selected to measure the corresponding concepts. Refer to the replication package for further information on the scales.

Security. To assess participants' security concerns with respect to the protection of their personal information, we used the 4-item security concern scale (i.e., "I worry about wrong information being linked to my identity due to security breaches") [1]. Responses were given on a 5-point answering scale ranging from strongly disagree (1) to strongly agree (5). Because results from the focus groups revealed two possible directions for the relationship between security concerns and self-hosting (i.e., providing more security and increasing security risks), we predicted a non-directional relationship between security concerns and self-hosting behavior.

Privacy. Participants' concerns regarding the availability of private information on the Internet were assessed with the 4-item privacy concern scale (i.e., "I am concerned that a person can find private information about me on the Internet") [16]. Responses were given on a 5-point answering scale ranging from strongly disagree (1) to strongly agree (5). Based on the results of the focus groups, we predicted a positive relation between privacy concerns and self-hosting behavior. This would indicate that self-hosting is accompanied by a higher concern for information privacy.

Technology interest and skills: Affinity for technology interaction (ATI). To measure participants' aptitude for interacting with technical systems, we used the 9-item ATI scale (i.e., "I try to understand how a technical system exactly works") [24]. Responses were given on a 6-point answering scale ranging from completely disagree (1) to completely agree (6). Based on the results of the focus groups, we predicted a positive relation between ATI and self-hosting behavior because self-hosters might show a higher interest in technical systems.

Openness to new things: Personal innovativeness in the domain of information technology (PIIT). We assessed participants' interest in trying out and experimenting with new technologies with the 4-item PIIT scale (i.e., "If I heard about a new information technology, I would look for ways to experiment with it") [2]. Responses were given on a 6-point answering scale ranging from strongly disagree (1) to strongly agree (6). Based on the results of the focus groups, we predicted a positive relation between PIIT and self-hosting behavior because self-hosters might be more open to trying out and experimenting with technologies.

Technology interest and skills: Computer self-efficacy. To assess participants' confidence in performing computer-related activities, we used the advanced (i.e., "Using a computer's task manager"; BITS-Ad) and expert ("Analyzing computer error log files"; BITS-Ex) subscales of the Brief Inventory of Technology Self-efficacy (BITS) [68], asking people to indicate their level of confidence performing each activity. Responses were given on a 6-point answering scale ranging from not at all confident (1) to completely confident (6). Based on the results of the focus groups, we predicted a positive relation between BITS-Ad and BITS-Ex and self-hosting behavior because self-hosters might show more advanced and expert technology skills.

Technology interest and skills: Self-hosting skills. Self-reported skills to set up and administrate a server were assessed with a self-developed 6-item scale. We first presented participants with a job description of a system administrator. Subsequently, we asked them to rate their abilities in different domains (computer networks, operating systems, servers [virtual or physical], software, system security, and system administration). Responses were given on a 6-point answering scale ranging from poor (1) to excellent (6). Based on the results of the focus groups, we predicted a positive relation between self-hosting skills and self-hosting behavior. This should be the case if self-hosting requires a certain basic skill set.

Technology interest and skills: IT background. To assess IT background, we used an item (i.e., "Are you studying or have you been working in any of the following areas: information, technology, computer science, electronic data processing, electrical engineering, communications technology, or similar?") introduced by Elbitar and colleagues [21]. Responses were given on a dichotomous answering scale (yes/no). Based on the results of the focus groups, we predicted a positive relation between IT background and self-hosting behavior. An IT background might provide people with the necessary skill set.

Tinkering and DIY: "Maker" activities. To measure how much time participants typically spend with domestic activities (e.g., baking), DIY activities (e.g., woodworking) and arts and crafts (e.g., ceramics), we adapted 18 different activities from Collier and Wayment [12] and asked participants to indicate their time spent with each activity on a scale from "none" (1) to "I spend large amount of time doing activity" (5). Based on the results of the focus groups, we predicted a positive relation between "maker" activities and self-hosting behavior. This should be the case if self-hosting goes along with other tinkering and DIY activities.

Tinkering and DIY: "Maker" self-identity. Participants' self-identity as a "maker" or DIY person was assessed by presenting them with a short description of what is meant by do-it-yourself (i.e., "sometimes this can be called crafting, sometimes it refers to hobbies. Typically it leads to making something tangible. That is, what you can create with your own two hands.") and asking them to indicate on a scale from not at all (1) to extremely so (5) how much they identify as a "maker" or DIY person. This procedure was adapted from Collier and Wayment [12]. Based on the results of the focus groups, we predicted a positive relation between "maker" self-identity and self-hosting behavior. This should be the case if self-hosters perceive themselves as DIY persons.

Money: Frugality. Participants' economical consumer lifestyle was assessed with the 9-item frugality scale (e.g., "I believe in being careful how I spend my money") [38]. Responses were given on a 6-point answering scale ranging from definitely disagree (1) to definitely agree (6). Based

Table 4: Consistency between operational and self-classification as self-hoster

	Self-classification		
	Non-Self-Hoster	Self-Hoster	Overall
Operational classification			
Non-Self-Hoster	355 (60.3%)	122 (20.7%)	477 (81.0%)
Self-Hoster	32 (5.4%)	80 (13.6%)	112 (19.0%)
Overall	387 (65.7%)	202 (34.3%)	589 (100.0%)

on the results of the focus groups, we predicted a positive relation between frugality and self-hosting behavior because self-hosters might ponder more about what to spend their money on.

Work effort: Grit. The extent of participants' consistency of interest (GRIT-Co) and perseverance of effort (GRIT-Pe) was assessed with the 8-item GRIT-S scale (e.g., "I often set a goal but later choose to pursue a different one"; "I am diligent.") [19]. Responses were given on a 5-point answering scale ranging from strongly disagree (1) to strongly agree (5). Based on the results of the focus groups, we predicted a positive relation between GRIT-Co and GRIT-Pe and self-hosting behavior. This might be the case if self-hosting behavior goes along with being consistent in an area of interest and putting effort into reaching a goal.

Control: Autonomy. To assess participants' valuation of self-direction and freedom of choice in their daily activities and undertakings, we adapted eight autonomy items (e.g., "I feel I'm doing what really interests me") from the Basic Psychological Need Satisfaction and Need Frustration Scale (BPNSNF) [10]. Responses were given on a 5-point answering scale ranging from not important to me at all (1) to very important to me (5). Based on the results of the focus groups, we predicted a positive relation between Autonomy and self-hosting behavior because self-hosting behavior might go along with the importance that is ascribed to having control over one's life.

7.1.3 Self-Report Identification

Because the identification of (non-)self-hosters in Survey 1 was based on operational criteria (i.e., participants' response behavior in closed and open questions), we employed a self-report identification procedure in Survey 2 to complement the measurement from the previous survey. To this end, we presented participants with a definition and examples of self-hosting and then asked them whether they would describe themselves as a self-hoster (i.e., whether they currently self-host or have recently self-hosted at least one tool/service in their personal life). We emphasized that their answer would not affect survey length or compensation to minimize externally motivated answering behavior.

7.2 Sample Selection Process

To allow for a meaningful comparison between self-hosters and non-self-hosters with reliable group estimates, self-hosters were matched with non-self-hosters in an approximate (influenced by the availability and responsiveness of participants in the pool) ratio of 1:3 [60] with respect to age, ethnicity, and sex, keeping the influence of these demographics as constant as possible in both groups. Choosing 3 non-self-hosters per 1 self-hoster, rather than 1:1, reduces error and confidence interval width when comparing the two groups [58]. To ensure estimates generalize to the broader U.S. population, our weighting scheme (Section 7.4) adjusts for non-proportional group sizes. For data collection, we again used Prolific. Data collection took place between September 2022 and November 2023. 98.68% of the sample was completed in December 2022. However, to increase the chances of including the maximum number of self-hosters, the survey was opened up to November 2023.

7.3 Data Cleaning and Preparation

We only included participants in the sample who had completed the entire survey and passed both attention checks. Participants who did not meet these criteria were excluded, and we re-sampled in line with the requirements of our sample selection process. Once data collection was completed, three researchers coded participants' final survey comments to see whether any comments raised doubts about data quality (e.g., participant indicated having trouble with filling out the scale items). In case of disagreement, they discussed their ratings and reached a consensus. Following this approach, one participant was excluded.

7.4 Survey Weighting

To be able to make valid claims about the population for our findings in Survey 2, we adjusted for deviations between our sample and the population with respect to the distribution of sex, age, and ethnicity as well as for the over-representation of self-hosters, both resulting from our sampling design (see section 7.2). The calibration weights are determined such that the

estimated weighted proportions across all socio-demographic groups containing self-hosters are the same in Survey 2 as they were in the representative Survey 1 and that the estimated share of self-hosters within each of these groups also corresponds to the prevalence estimated in Survey 1 (see tables 7 and 8). As in section 4.4, we used the Generalized Regression estimator [14] for this purpose. The weights for Survey 2 are therefore $w_h = \widehat{N}_h^{(1)} \Big/ n_h^{(2)}$ for each element in the h-th group defined by sex, age, ethnicity, and self-hoster status. $\widehat{N}_h^{(1)}$ denotes the *weighted* size of this group in Survey 1 and $n_h^{(2)}$ the sample size in Survey 2.

8 Characteristics Survey 2 - Results

Our initial sample in Survey 2 consisted of $n = 112$ self-hosters (i.e., a retention rate of 90.32%) and $n = 477$ non-self-hosters (self-selected from a pool of $n = 1355$), identified in Survey 1. Appendix B contains details on sample demographics.

Because we asked participants in Survey 2 to indicate whether they described themselves as self-hoster, we used this indication to compare it with the classification based on operational criteria from Survey 1. Table 4 shows the consistency between classifications of Survey 1 and Survey 2. The results reported hereafter follow a conservative approach and are thus solely based on participants whose classification concurred.

8.1 RQ3: Individual Characteristics

We inspected all scales and, if applicable, subscales, with respect to their internal consistencies [33]. We calculated Cronbach's alpha (α) for all scales requiring simple mean scores [1, 2, 12, 21, 24, 38, 68] and McDonald's omega (ω) for all scales requiring mean scores weighted with factor loadings [10, 16, 19]. Internal consistency ranged from $\alpha = .82$ to .95 and $\omega = .77$ to .92, indicating an overall acceptable to good reliability for the scales. To investigate which of our selected predictors best explain self-hosting behavior, we entered all predictors into a backward stepwise regression analysis (see Table 5). The model that best fits our data, as indicated by the lowest Akaike information criterion (AIC) [3], contains affinity for technology interaction (ATI), "maker" self-identity (DIY-self), perseverance of effort (GRIT-Pe), IT background and self-reported self-hosting skills as significant predictors. ATI, DIY-self, IT background, and skills showed significant positive relations to self-hosting behavior, indicating that participants who belong to the group of self-hosters show a higher aptitude for interacting with technical systems, identify themselves stronger as makers, report more frequently having an IT background and report better self-hosting skills than participants who belong to the group of non-self-hosters. GRIT-Pe showed significant negative relations with self-hosting behavior, indicating that self-hosters report less perseverance in their efforts as compared to non-self-hosters.

9 Characteristics Survey 2 - Discussion

The goal of Survey 2 was to get a better understanding of self-hosters with respect to their individual characteristics. Our results showed, in line with our predictions that self-hosters (as compared to non-self-hosters) show greater interest in technical systems, more often have a skill set that allows them to perform the behavior, and more frequently have an IT background. In Survey 1, we learned that self-hosters in general use more tools, including SaaS solutions. Both findings suggest that self-hosting goes hand in hand with a strong technical background. Speculatively, technical people who use a broad set of tools also adopt self-hostable solutions. Our observations may also point to major roadblocks in the self-hosting ecosystem, allowing only skilled people to stay in it.

Moreover, we found that self-hosters perceive themselves more often as DIY persons, although we did not find evidence that self-hosters' DIY activities differ from those of non-self-hosters. Contrary to our prediction, self-hosters seem to show less perseverance in their efforts as compared to non-self-hosters. What might appear as a counter-intuitive finding at first glance could be explained by research showing that having grit not only helps people to achieve difficult goals [18] but can also have a flip side, making it hard for people to let go [4], therefore making them persist when moving on might be the better choice [42]. Accordingly, the fact that self-hosters show less perseverance of effort also indicates that they might have an easier time letting go of goals that are not worth pursuing. Further research should explore whether and how being more flexible in goal pursuit might aid or result from self-hosting behavior.

Unexpectedly, we did not find any evidence that self-hosters differ from non-self-hosters with respect to their security or privacy concerns, their computer self-efficacy, their openness to new technologies, their economical consumer lifestyle, or their valuation of autonomy. This is especially surprising because self-hosters named privacy, autonomy, and security as motivational factors [26]. However, our results do not imply that security or privacy concerns play no role when it comes to self-hosting. Rather, these factors do not explain the behavior beyond the predictors discussed above.

At present, we can only speculate that, although carefully considered, selected scale measures might not exactly represent the core themes identified in the focus groups (e.g., did the frugality scale capture all financial aspects involved in self-hosting?) or identified core themes might not apply to all self-hosters but potentially only to a specific subgroup (i.e., concurring themes in Gröber et al.'s research are based on a specific community, that is Nextcloud users [26]). It is

Table 5: Logistic regression model for self-hoster status: stepwise selection

	Model 00	Model 01	Model 02	Model 03	Model 04	Model 05	Model 06	Model 07	Model 08	Model 09
Intercept	0.1	0.6	0.6	−0.2	−0.3	−0.9	−0.8	−0.6	0.3	0.5
ATI	0.3	0.2	0.5	0.4	0.4	0.7 **	0.7 **	0.7 **	0.7 **	0.7 **
DIY-self	0.5 *	0.5 *	0.5 *	0.5 *	0.5 *	0.6 **	0.6 **	0.6 **	0.5 **	0.5 **
GRIT-pe	−1.5 ***	−1.6 ***	−1.5 ***	−1.6 ***	−1.6 ***	−1.4 ***	−1.7 ***	−1.7 ***	−1.8 ***	−1.8 ***
IT background	0.9 *	0.9 *	0.8 *	0.8 *	0.9 *	1.0 **	1.0 **	1.0 **	0.9 *	0.9 *
Skills	1.1 ***	1.1 ***	1.1 ***	1.1 ***	1.2 ***	1.4 ***	1.4 ***	1.4 ***	1.4 ***	1.4 ***
Privacy	0.2	0.2	0.2	0.1	0.2	0.2	0.2	0.1	0.1	
DIY activities	−0.9	−0.9	−0.8	−0.8	−0.8	−0.9	−1.0	−1.0		
Security	0.0	−0.1	0.0	−0.1	−0.1	−0.1	−0.1			
GRIT-co	−0.4	−0.3	−0.4	−0.4	−0.4	−0.4				
BITS-ad	1.0 *	1.0 *	1.0 *	1.0 *	1.0 **					
BITS-ex	0.1	0.1	0.1	0.1						
Autonomy	−0.3	−0.3	−0.2							
PIIT	0.3	0.3								
Frugality	−0.7									
Deviance	131.4	132.7	133.6	133.9	134.0	138.3	140.0	140.1	141.4	141.4
AIC	167.3	164.6	161.8	159.3	157.0	154.7	152.8	151.0	149.7	148.2
# of observ.	432	432	432	432	432	432	432	432	432	432

$^{***}p < 0.001$ $^{**}p < 0.01$ $^{*}p < 0.05$
Results after controlling for the effects of Gender, Age and Ethnicity

also possible that identified core themes are at least partly influenced by focus group participants' ideas and conceptions about self-hosters and that these conceptions do not perfectly match the actual characteristics of self-hosters.

10 Discussion and Future Research Directions

In this section, we discuss future research directions and highlight opportunities and challenges.

Security and Usability of Infrastructure. This work focuses on people who are currently self-hosting. Thus, we must assume survivorship bias with regard to the perspective of people who would like to become self-hosters or people who tried and failed. Based on our findings, having technical skills (or believing to have technical skills) and IT background are indicators for self-hosting. This possession of technical skills could lead to people "surviving" when self-hosting. However, it may actually indicate severe technical roadblocks or usability challenges. Thus, future research should investigate what is currently preventing people from self-hosting. In doing so, studies should maximize external validity to provide a realistic view of entry burdens such as set-up procedures, infrastructure decisions, and secure configurations. We argue further that research may focus on usability and security challenges of hosting infrastructure in the long run. In general, conducting research tailored to assist private hosters, who may have fewer resources and background knowledge, will benefit the greater population of IT administrators if security and usability challenges are streamlined.

Investigate Socio-Technical Influences. Infrastructure does not exist in a vacuum but is directly impacted by the people administrating it and the social environment they are embedded in [26]. We identified different individual characteristics that (may) predict self-hosting behavior. Future research should investigate the interplay of individual characteristics enabling or constraining different stages of the hosting process. The social embeddedness may actually be a determining factor of self-hosting success. Long-term studies could help to link infrastructure configurations with socio-technical parameters, such as individual characteristics of administrators, and relate those to observable security outcomes. This way we obtain valid assessments of the security of self-hosted systems. Moreover, this work identifies demographic data that describes the self-hosting community. This data identifies prominent populations of self-hosters and highlights underrepresented groups. For example, we found that gender minorities are less likely to self-host. Future research might investigate this beyond organizational embedded administrators [34].

Community-Driven Design. While we cannot directly offer implications for design based on our findings, the demographic data we collected provide valuable perspectives for future design efforts. Specifically, it is useful for building personas representative of the self-hosting population, while being mindful of underrepresented groups. We suggest community-driven and participatory methodologies when designing solutions or tooling for self-hosters. That is because of the various use cases and diverse demographical traits, which makes one-fits-all solutions unlikely.

11 Ethical Considerations

Both studies received approval by Saarland University's ethical review board. Before collecting any data, we obtained informed consent of the participants for both studies. We told the participants that the survey was anonymous and that all data will be treated confidentially. Moreover, we clarified that their participation was voluntary and that they had the right to withdraw at any point. We disclosed our identity and offered a contact email address for any questions. We were transparent about the overall study process including an optional follow-up survey. The collected data was stored on CISPA Helmholtz Center for Information Security private servers. To protect the participants' privacy, we anonymized the study data we made available to the public.

12 Limitations

Generalizability. With our sampling and weighting method, we recruited a representative sample of the United States with respect to age, sex, and ethnicity. Our sample might still be biased by our approach to recruiting participants only via Prolific. However, recent research showed that sampling on Prolific does allow to generalize results at least with respect to certain topics [64] and outperforms other means of online data collection [17,49]. Accordingly, in terms of overall feasibility, our approach maximizes the currently available resources and instruments.

Self-selection Bias. Our sampling method and, thus, our results are not immune against self-selection bias [29]. However, we took utmost care when announcing the survey and creating the instructions to conceal the actual purpose of our research. Accordingly, we cannot entirely rule out self-selection bias due to the general topic (e.g., software and application usage), but we minimized self-selection based on the topic of self-hosting.

Framing of Self-Hosting. The results of our research largely rest on our definition of self-hosting, which also determined the selection of use cases and tools. Accordingly, our research might underestimate other instances in which people administrate their own infrastructure and services but did not see their behavior reflected in our definition of self-hosting. We paid close attention while coding the open responses in Survey 1 to include all possible cases that went beyond our pre-selection of use cases (e.g., gaming, media server) and are confident to have included these in our classification. Yet, the self-report identification in Survey 2 (i.e., whether people would describe themselves as self-hoster, taking our definition into account) might have led to an exclusion of actual self-hosters who do not agree with our definition. Because it is not possible to conduct the present research without at least a working definition of self-hosting, future research should explore other use cases and potentially more inclusive or more narrow sub-definitions of self-hosting. For example, we included smart home devices as an edge-case if people administrated a home automation service in a server-client model. Future work could look into sub-aspects of self-hosting that only satisfy parts of our definition like "responsibility", "available over a network", and "software control", e.g., for open-source IoT devices.

Causality. As our results rest on correlations, we cannot make any claims about causal relationships between self-hosting behavior and individual characteristics [41]. More specifically, we do not know whether interest in technical systems, self-hosting skills, and having an IT background is a precondition for self-hosting or follows from self-hosting behavior. Likewise, perceiving oneself as a DIY person might be a necessary prerequisite or might simply result from the experience of self-hosting. Similarly, having an easier time letting go of goals might be beneficial for self-hosting in a fast-moving technology ecosystem or might be a result of having experienced the need to adapt and shift goals quickly when practicing self-hosting.

13 Conclusion

We find that self-hosting is not a niche phenomenon, with an upper-bound estimate of 8.4% of the U.S. population. Results identify prevalent potential high-risk use cases such as Wordpress instances running on home servers and home automation accessible from the Internet. Our work lays the foundations for further research about a population previously assumed to be strongly motivated by security and privacy concerns [26]. However, we find that people who have higher privacy or security concerns are just as likely to be self-hosting as those who do not have such concerns. In other words: neither privacy nor security concerns are predictors of self-hosting behavior. Instead, we find that characteristics related to technology interest, hosting skills, and "maker" self-identity are positively correlated with self-hosting. Since we only consider present self-hosters, these characteristics might be subject to survivorship bias. Our findings may actually indicate severe technical roadblocks or usability challenges in the hosting ecosystem. Thus, our findings can inform the design of new solutions that benefit both expert and novice administrators.

Acknowledgments

We thank the reviewers and the shepherd of this paper for their time, valuable feedback, and guidance, which helped to improve our work. We also thank the participants of the focus groups and surveys for their important contributions and insights. Additionally, we thank our colleagues Matthias Fassl, Sebastian Roth, and Marius Steffens for their helpful feedback and support throughout this project. Your efforts and encouragement have been greatly appreciated.

References

[1] Joyce Hoese Addae, Michael Brown, Xu Sun, Dave Towey, and Milena Radenkovic. Measuring attitude towards personal data for adaptive cybersecurity. *Information & Computer Security*, 25(5):560–579, 2017.

[2] Ritu Agarwal and Jayesh Prasad. A conceptual and operational definition of personal innovativeness in the domain of information technology. *Information systems research*, 9(2):204–215, 1998.

[3] Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akadémiai Kiado.

[4] Larbi Alaoui and Christian Fons-Rosen. Know when to fold'em: The flip side of grit. *European Economic Review*, 136:103736, 2021.

[5] Philip Banyard, Andrew Grayson, and MT Orne. Demand characteristics. *Introducing psychological research: Sixty studies that shape psychology*, pages 395–401, 1996.

[6] Sara Bly, Bill Schilit, David W McDonald, Barbara Rosario, and Ylian Saint-Hilaire. Broken expectations in the digital home. In *CHI'06 extended abstracts on Human factors in computing systems*, 2006.

[7] Hudson Borges, Andre Hora, and Marco Tulio Valente. Understanding the factors that impact the popularity of github repositories. In *2016 IEEE international conference on software maintenance and evolution (ICSME)*, pages 334–344. IEEE, 2016.

[8] AJ Brush. It@ home: Often best left to professionals. In *Position paper for the CHI 2006 Workshop on IT@ Home*, 2006.

[9] Kathy C. Charmaz. *Constructing grounded theory*. Sage, 2014.

[10] Beiwen Chen, Maarten Vansteenkiste, Wim Beyers, Liesbet Boone, Edward L Deci, Jolene Van der Kaap-Deeder, Bart Duriez, Willy Lens, Lennia Matos, Athanasios Mouratidis, et al. Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motivation and emotion*, 39:216–236, 2015.

[11] J Chen, RR Sitter, and C Wu. Using Empirical Likelihood Methods to Obtain Range Restricted Weights in Regression Estimators for Surveys. *Biometrika*, 89(1):230–237, 2002.

[12] Ann Futterman Collier and Heidi A Wayment. Psychological benefits of the "maker" or do-it-yourself movement in young adults: A pathway towards subjective well-being. *Journal of Happiness Studies*, 19:1217–1239, 2018.

[13] Juliet M. Corbin and Anselm L. Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, 19(6):418–427, 1990.

[14] Jean-Claude Deville and Carl-Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992.

[15] Constanze Dietrich, Katharina Krombholz, Kevin Borgolte, and Tobias Fiebig. Investigating system operators' perspective on security misconfigurations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1272–1289, 2018.

[16] Tamara Dinev and Paul Hart. An extended privacy calculus model for e-commerce transactions. *Information systems research*, 17(1):61–80, 2006.

[17] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720, 2023.

[18] Angela L Duckworth, Christopher Peterson, Michael D Matthews, and Dennis R Kelly. Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*, 92(6):1087, 2007.

[19] Angela Lee Duckworth and Patrick D Quinn. Development and validation of the short grit scale (grit–s). *Journal of personality assessment*, 91(2):166–174, 2009.

[20] W Keith Edwards and Rebecca E Grinter. At home with ubiquitous computing: Seven challenges. In *Ubicomp: Ubiquitous Computing: International Conference*, 2001.

[21] Yusra Elbitar, Michael Schilling, Trung Tin Nguyen, Michael Backes, and Sven Bugiel. Explanation beats context: The effect of timing & rationales on users' runtime permission decisions. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 785–802, 2021.

[22] Sascha Fahl, Yasemin Acar, Henning Perl, and Matthew Smith. Why eve and mallory (also) love webmasters: A study on the root causes of ssl misconfigurations. In *Proceedings of the 9th ACM symposium on Information, computer and communications security (CCS)*, pages 507–512, 2014.

[23] Tobias Fiebig, Seda Gürses, Carlos H Gañán, Erna Kotkamp, Fernando Kuipers, Martina Lindorfer,

Menghua Prisse, and Taritha Sari. Heads in the clouds: Measuring the implications of universities migrating to public clouds. *The 23rd Privacy Enhancing Technologies Symposium*, 2023.

[24] Thomas Franke, Christiane Attig, and Daniel Wessel. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ati) scale. *International Journal of Human–Computer Interaction*, 35(6):456–467, 2019.

[25] Rebecca E Grinter, W Keith Edwards, Mark W Newman, and Nicolas Ducheneaut. The work to make a home network work. In *ECSCW: European Conference on Computer-Supported Cooperative Work*, 2005.

[26] Lea Gröber, Rafael Mrowczynski, Nimisha Vijay, Daphne A Muller, Adrian Dabrowski, and Katharina Krombholz. To cloud or not to cloud: A qualitative study on self-hosters' motivation, operation, and security mindset. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2491–2508, 2023.

[27] Junxiao Han, Shuiguang Deng, Xin Xia, Dongjing Wang, and Jianwei Yin. Characterization and prediction of popular projects on github. In *2019 IEEE 43rd annual computer software and applications conference (COMPSAC)*, volume 1, pages 21–26. IEEE, 2019.

[28] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of "big data" on cloud computing: Review and open research issues. *Information systems*, 47:98–115, 2015.

[29] James J Heckman. Selection bias and self-selection. In *Econometrics*, pages 201–224. Springer, 1990.

[30] Franziska Herbert, Steffen Becker, Leonie Schaewitz, Jonas Hielscher, Marvin Kowalewski, Angela Sasse, Yasemin Acar, and Markus Dürmuth. A world full of privacy and security (mis) conceptions? findings of a representative survey in 12 countries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2023.

[31] Alicia Hope. Toyota connected service decade-long data leak exposed 2.15 million customers. CPO Magazine, https://www.cpomagazine.com/cyber-security/toyota-connected-service-decade-long-data-leak-exposed-2-15-million-customers/, last accessed: 2/9/2023, 2023.

[32] Wenjin Hu, Tao Yang, and Jeanna N Matthews. The good, the bad and the ugly of consumer cloud storage. *ACM SIGOPS Operating Systems Review*, 44(3), 2010.

[33] Michael T Kalkbrenner. Alpha, omega, and h internal consistency reliability estimates: Reviewing these options and when to use them. *Counseling Outcome Research and Evaluation*, 14(1):77–88, 2023.

[34] Mannat Kaur, Harshini Sri Ramulu, Yasemin Acar, and Tobias Fiebig. "oh yes! over-preparing for meetings is my jam:)": The gendered experiences of system administrators. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

[35] Sara Kraemer and Pascale Carayon. Human errors and violations in computer and information security: The viewpoint of network administrators and security specialists. *Applied ergonomics*, 38(2):143–154, 2007.

[36] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel Von Zezschwitz. ïf https were secure, i wouldn't need 2fa¨ – end user and administrator mental models of https. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 246–263. IEEE, 2019.

[37] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar Weippl. " i have no idea what i'm doing"-on the usability of deploying {HTTPS}. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 1339–1356, 2017.

[38] John L Lastovicka, Lance A Bettencourt, Renee Shaw Hughner, and Ronald J Kuntze. Lifestyle of the tight and frugal: Theory and measurement. *Journal of consumer research*, 26(1):85–98, 1999.

[39] Sunghee Lee and Richard Valliant. Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods & Research*, 37(3):319–343, 2009.

[40] Frank Li, Lisa Rogers, Arunesh Mathur, Nathan Malkin, and Marshini Chetty. Keepers of the machines: Examining how system administrators manage software updates for multiple machines. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 273–288, 2019.

[41] Scott O Lilienfeld. Correlation still isn't causation. *APS Observer*, 19, 2006.

[42] Gale M Lucas, Jonathan Gratch, Lin Cheng, and Stacy Marsella. When the going gets tough: Grit predicts costly perseverance. *Journal of Research in Personality*, 59:15–22, 2015.

[43] Michael Martin. Computer and Internet Use in the United States: 2018. American Community Survey Reports – ACS-49, 2021.

[44] Peter Mell and Tim Grance. The nist definition of cloud computing. *NIST Special Publication 800-145*, 2011.

[45] David Molnar and Stuart E Schechter. Self hosting vs. cloud hosting: Accounting for the security impact of hosting in the cloud. In *WEIS*, 2010.

[46] Julian Oliver. 36c3 - server infrastructure for global rebellion. Youtube, https://www.youtube.com/watch?v=I_O3zj3p52A, last accessed: 2/9/2023, 2019.

[47] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, 45(4):867–872, 2009.

[48] Carly Page. Microsoft ai researchers accidentally exposed terabytes of internal sensitive data. TechCrunch, https://techcrunch.com/2023/09/18/microsoft-ai-researchers-accidentally-exposed-terabytes-of-internal-sensitive-data/, last accessed: 2/9/2023, 2023.

[49] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, page 1, 2022.

[50] Danny Pfeffermann. Methodological Issues and Challenges in the Production of Official Statistics: 24th Annual Morris Hansen Lecture. *Journal of Survey Statistics and Methodology*, 3(4):425–483, 2015.

[51] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How i learned to be secure: a census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 666–677, 2016.

[52] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. Where is the digital divide? a survey of security, privacy, and socioeconomics. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 931–936, 2017.

[53] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1326–1343. IEEE, 2019.

[54] Benjamin M. Reist and Richard Valliant. Model-assisted estimators for time-to-event data from complex surveys. *Statistics in Medicine*, 39(29):4351–4371.

[55] Melanie A Revilla and Willem E Saris. A comparison of the quality of questions in a face-to-face and a web survey. *International Journal of Public Opinion Research*, 25(2):242–253, 2013.

[56] Thea Riebe, Tom Biselli, Marc-André Kaufhold, and Christian Reuter. Privacy concerns and acceptance factors of osint for cybersecurity: A representative survey. *Proceedings on Privacy Enhancing Technologies*, (1):477–493, 2023.

[57] Camille Ryan. Computer and Internet Use in the United States: 2016. 2018.

[58] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

[59] Anselm L. Strauss and Juliet M. Corbin. *Grounded theory in practice*. Sage, 1997.

[60] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[61] Jayachander Surbiryala and Chunming Rong. Cloud computing: History and overview. In *2019 IEEE Cloud Summit*. IEEE, 2019.

[62] Dan Svantesson and Roger Clarke. Privacy and consumer risks in cloud computing. *Computer law & security review*, 26(4):391–397, 2010.

[63] Nestori Syynimaa and Tessa Viitanen. Is my office 365 GDPR compliant?: Security issues in authentication and administration. In *International Conference on Enterprise Information Systems*, 2018.

[64] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How well do my results generalize now? the external validity of online privacy and security surveys. In *Eighteenth symposium on usable privacy and security (SOUPS 2022)*, pages 367–385, 2022.

[65] Peter Tolmie, Andy Crabtree, Tom Rodden, Chris Greenhalgh, and Steve Benford. Making the home network at home: Digital housekeeping. In *European Conference on Computer-Supported Cooperative Work*, 2007.

[66] prolific.com. How does prolific create the demographic subgroups used for representative samples? https://researcher-help.prolific.com/hc/en-gb/articles/360019238413-Representative-samples-FAQ#heading-2, 2022. [Accessed: 2024/02/06].

[67] U.S. Census Bureau. Decennial census 2010, October 2022.

[68] Arne Weigold and Ingrid K Weigold. Measuring confidence engaging in computer activities at different skill levels: Development and validation of the brief inventory of technology self-efficacy (bits). *Computers & Education*, 169:104210, 2021.

[69] Dominik Wermke, Nicolas Huaman, Christian Stransky, Niklas Busch, Yasemin Acar, and Sascha Fahl. Cloudy with a chance of misconceptions: exploring users' perceptions and expectations of security and privacy in cloud office suites. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 359–377, 2020.

[70] Zeljka Zorz. A bug revealed chatgpt users' chat history, personal and billing data. Help Net Security, https://www.helpnetsecurity.com/2023/03/27/chatgpt-data-leak/, last accessed: 2/9/2023, 2023.

A Large-Scale Studies on Technical Topics

Identifying a Sub-Population Based on Operational Criteria and Self-Report. One of the main goals of our research was to identify a group of people based on their specific behavior (i.e., private self-hosting). We spent time and effort to (1) define the behavior of interest as exactly as possible, and (2) derive measurable indicators of usage from this definition. This allowed us to use operational criteria for identification without the necessity for participants to self-identify. Using operational criteria can be advantageous to avoid answering behavior based on demand characteristics [5] and to ensure that the behavior intended to be captured is represented entirely in what is actually measured. However, this method is not immune to participants' misconceptions influencing their response behavior. To illustrate, without relevant background knowledge, downloading and installing an application might be misconceived as administrating it on their own hardware. Accordingly, the right wording of items and questions is paramount in minimizing the number of false positives for identification. Yet, exact wording cannot entirely rule out misconceptions.

A remedy against misconceptions is educating participants about the behavior of interest. Under the premise that participants read the information attentively, understand the definition, and can apply it to their own behavior, false positives in identification might be reduced. Yet, such an approach is more susceptible to demand characteristics. In our research, we combined both methods to keep participants' misconceptions and demand characteristics at a minimum. We suggest that this might be good practice for investigating certain behaviors with surveys to balance out trade-offs.

Representative Sampling and Data Processing. In order to assess how commonly individuals host their own services and understand how this behavior relates to personal traits, an important task was to facilitate drawing conclusions from volunteer web surveys (i.e., Prolific) that are reasonably generalizable to the population. As part of our two web surveys, we put a lot of effort into increasing the precision of our results by (1) carefully selecting a sample that reflects the overall population in terms of age, gender, and ethnicity, and (2) mitigating potential biases and selectivity in responses through thorough data cleaning and the application of calibration weighting.

We are convinced that such efforts to capture the population's full diversity in the sample and minimize the impact of potential selectivity and bias (e.g., due to low-quality responses) contribute significantly to advancing human subjects research in the area of security and privacy.

B Sample Composition and Self-Hosting Prevalence by Socio-Demographic Groups

Table 6: Share of sex × age × ethnicity groups in our second survey (in %)

Sex × Age	Asian	Black	Mixed	Other	White	Overall
Female						
18 – 28	0.7	1.1	0.0	0.0	2.1	3.9
28 – 38	0.0	2.9	0.0	0.0	1.8	4.7
38 – 48	0.7	1.1	0.0	0.0	4.9	6.7
48 – 58	0.0	0.7	0.0	0.0	4.4	5.0
58 – 150	0.0	0.0	0.0	0.0	4.2	4.2
Overall	1.3	5.9	0.0	0.0	17.4	24.6
Male						
18 – 28	0.2	0.7	0.7	0.2	9.6	11.2
28 – 38	1.1	2.3	0.0	1.8	9.8	15.0
38 – 48	0.7	1.0	0.3	1.6	13.8	17.4
48 – 58	0.7	1.8	0.0	1.0	15.4	18.9
58 – 150	0.0	2.0	0.0	0.7	10.4	13.0
Overall	2.6	7.6	1.0	5.2	59.0	75.4
Overall						
18 – 28	0.8	1.8	0.7	0.2	11.7	15.1
28 – 38	1.1	5.2	0.0	1.8	11.5	19.7
38 – 48	1.3	2.1	0.3	1.6	18.7	24.1
48 – 58	0.7	2.4	0.0	1.0	19.8	23.9
58 – 150	0.0	2.0	0.0	0.7	14.6	17.2
Overall	3.9	13.5	1.0	5.2	76.4	100.0

Table 7: Share of sex × age × ethnicity groups in the population and our survey (in %)

Sex × Age	Asian		Black		Mixed		Other		White		Overall	
	Survey	Population	Survey	Population	Survey	Population	Survey	Population	Survey	Population	Survey	Population
Female												
18 – 28	0.5	0.5	1.3	1.3	0.3	0.3	0.9	0.8	6.0	6.1	9.0	9.1
28 – 38	0.6	0.6	1.2	1.2	0.3	0.2	0.8	0.8	5.8	5.8	8.6	8.5
38 – 48	0.5	0.5	1.1	1.2	0.2	0.2	0.7	0.6	6.6	6.6	9.1	9.2
48 – 58	0.5	0.4	1.2	1.2	0.1	0.2	0.5	0.4	7.4	7.3	9.7	9.5
58+	0.6	0.6	1.5	1.5	0.2	0.2	0.3	0.4	12.4	12.5	15.0	15.1
Overall	2.7	2.6	6.3	6.4	1.1	1.1	3.1	3.1	38.2	38.3	51.4	51.5
Male												
18 – 28	0.5	0.5	1.3	1.3	0.3	0.3	1.0	1.0	6.4	6.3	9.5	9.4
28 – 38	0.5	0.5	1.1	1.1	0.2	0.2	0.9	0.9	5.8	5.9	8.6	8.6
38 – 48	0.5	0.5	1.1	1.1	0.2	0.2	0.7	0.7	6.6	6.6	9.0	9.1
48 – 58	0.4	0.4	1.0	1.0	0.1	0.1	0.5	0.5	7.0	7.1	9.0	9.1
58+	0.5	0.4	1.1	1.1	0.1	0.1	0.3	0.3	10.4	10.4	12.4	12.4
Overall	2.3	2.3	5.6	5.6	1.0	1.0	3.3	3.3	36.3	36.4	48.6	48.5
Overall												
18 – 28	0.9	1.0	2.7	2.6	0.7	0.6	1.9	1.8	12.4	12.4	18.5	18.5
28 – 38	1.1	1.1	2.3	2.2	0.5	0.5	1.7	1.7	11.6	11.6	17.2	17.1
38 – 48	1.0	1.0	2.2	2.3	0.4	0.4	1.3	1.3	13.2	13.3	18.1	18.3
48 – 58	0.9	0.8	2.2	2.2	0.3	0.3	0.9	0.9	14.5	14.4	18.7	18.6
58+	1.1	1.0	2.5	2.5	0.3	0.3	0.7	0.7	22.9	22.9	27.4	27.5
Overall	5.0	4.9	11.9	12.0	2.1	2.1	6.4	6.4	74.6	74.7	100.0	100.0

Table 8: Estimated self-hosting prevalence by sex, age and ethnicity (in %)

Sex × Age	Asian	Black	Mixed	Other	White	Overall
female						
18 – 28	0.0	5.0 ± 9.8	0.0	0.0	2.2 ± 3.1	2.2 ± 2.5
28 – 38	0.0	12.5 ± 16.7	0.0	0.0	3.4 ± 3.9	3.9 ± 3.4
38 – 48	12.5 ± 24.5	0.0	0.0	0.0	6.1 ± 4.8	5.1 ± 3.7
48 – 58	0.0	0.0	0.0	14.3 ± 28.0	2.7 ± 3.0	2.7 ± 2.6
58+	0.0	0.0	0.0	0.0	1.6 ± 1.8	1.3 ± 1.5
Overall	2.6 ± 5.0	3.2 ± 3.5	0.0	2.0 ± 4.0	3.0 ± 1.4	2.9 ± 1.2
male						
18 – 28	14.3 ± 28.0	5.0 ± 9.8	40.0 ± 48.0	0.0	17.2 ± 7.7	14.3 ± 5.7
28 – 38	37.5 ± 35.9	41.2 ± 24.1	0.0	23.1 ± 23.8	11.9 ± 7.0	18.1 ± 6.6
38 – 48	14.3 ± 28.0	0.0	0.0	11.1 ± 21.8	12.6 ± 6.7	10.8 ± 5.4
48 – 58	0.0	26.7 ± 23.2	0.0	16.7 ± 32.7	22.1 ± 8.0	21.1 ± 7.0
58+	0.0	20.0 ± 21.0	0.0	20.0 ± 39.2	8.4 ± 4.4	9.3 ± 4.2
Overall	14.3 ± 11.5	18.0 ± 7.9	12.3 ± 14.7	12.6 ± 9.7	14.0 ± 2.9	14.3 ± 2.5
Overall						
18 – 28	7.1 ± 14.0	5.0 ± 6.9	19.7 ± 23.6	0.0	9.7 ± 4.2	8.3 ± 3.2
28 – 38	17.5 ± 16.8	26.9 ± 14.7	0.0	12.2 ± 12.6	7.6 ± 4.0	10.9 ± 3.7
38 – 48	13.3 ± 18.5	0.0	0.0	5.5 ± 10.7	9.3 ± 4.1	7.9 ± 3.2
48 – 58	0.0	12.5 ± 10.8	0.0	15.4 ± 21.3	12.2 ± 4.2	11.6 ± 3.6
58+	0.0	8.0 ± 8.4	0.0	9.2 ± 18.1	4.7 ± 2.2	4.9 ± 2.1
Overall	8.0 ± 6.0	10.0 ± 4.1	5.9 ± 7.1	7.4 ± 5.3	8.3 ± 1.6	8.4 ± 1.4

± indicates the lower and upper bounds of the 95% confidence intervals