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 ($n = 1505$). (2) We run a follow-up survey on Prolific ($n = 589$) 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-
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

\(^1\)https://projects.cispa.saarland/lea.groeber/usenix24-sh-prevalence
behavior 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. 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.

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 crowdsourced 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.

4.3 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). Yet, 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 \(w_h = n_h/n_h^{(1)}\) 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>
<thead>
<tr>
<th>Prevalence</th>
<th>Distribution in subgroups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>18 – 28</td>
<td>8.3 ± 3.2</td>
</tr>
<tr>
<td>28 – 38</td>
<td>10.9 ± 3.7</td>
</tr>
<tr>
<td>38 – 48</td>
<td>7.9 ± 3.2</td>
</tr>
<tr>
<td>48 – 58</td>
<td>11.6 ± 3.6</td>
</tr>
<tr>
<td>58</td>
<td>4.9 ± 2.1</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
</tr>
<tr>
<td>asian</td>
<td>8.0 ± 6.0</td>
</tr>
<tr>
<td>black</td>
<td>10.0 ± 4.1</td>
</tr>
<tr>
<td>mixed</td>
<td>5.9 ± 7.1</td>
</tr>
<tr>
<td>other</td>
<td>7.4 ± 5.3</td>
</tr>
<tr>
<td>white</td>
<td>8.3 ± 1.6</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>2.9 ± 1.2</td>
</tr>
<tr>
<td>male</td>
<td>14.3 ± 2.5</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>woman</td>
<td>3.1 ± 1.2</td>
</tr>
<tr>
<td>man</td>
<td>14.1 ± 2.6</td>
</tr>
<tr>
<td>non-binary</td>
<td>0.0</td>
</tr>
<tr>
<td>self-described</td>
<td>11.0 ± 20.3</td>
</tr>
<tr>
<td>not stated</td>
<td>18.5 ± 33.1</td>
</tr>
</tbody>
</table>

Bolded 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 RQ1: 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-
To better understand self-hosting behavior, we looked at the hosters solely based on their responses in other open response summarized as 'Other'. Participants who we identified as self-hosters. Tools additionally listed by these participants under distribution of the different use cases and tools selected from 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

<table>
<thead>
<tr>
<th>Tool</th>
<th>Self-Hosted by</th>
<th>Server type</th>
<th>Other / unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Home</td>
<td>VPS</td>
</tr>
<tr>
<td>Communication</td>
<td></td>
<td>20.2</td>
<td>64.5</td>
</tr>
<tr>
<td>Teamspeak</td>
<td></td>
<td>12.0</td>
<td>63.3</td>
</tr>
<tr>
<td>Mumble</td>
<td></td>
<td>8.0</td>
<td>32.0</td>
</tr>
<tr>
<td>Rocket.Chat</td>
<td></td>
<td>2.7</td>
<td>65.8</td>
</tr>
<tr>
<td>Jitsi Meet</td>
<td></td>
<td>2.6</td>
<td>32.1</td>
</tr>
<tr>
<td>Mattermost</td>
<td></td>
<td>1.3</td>
<td>31.6</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>1.4</td>
<td>0.0</td>
</tr>
<tr>
<td>File Storage</td>
<td></td>
<td>17.7</td>
<td>74.9</td>
</tr>
<tr>
<td>Nextcloud</td>
<td></td>
<td>4.1</td>
<td>49.0</td>
</tr>
<tr>
<td>ownCloud</td>
<td></td>
<td>2.8</td>
<td>100.0</td>
</tr>
<tr>
<td>SparkleShare</td>
<td></td>
<td>2.6</td>
<td>65.7</td>
</tr>
<tr>
<td>Synthing</td>
<td></td>
<td>1.4</td>
<td>100.0</td>
</tr>
<tr>
<td>Seafile</td>
<td></td>
<td>6.7</td>
<td>59.5</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>16.3</td>
<td>73.3</td>
</tr>
<tr>
<td>Synchronized PW Managing</td>
<td></td>
<td>6.8</td>
<td>33.0</td>
</tr>
<tr>
<td>Vault-bitwarden</td>
<td></td>
<td>2.7</td>
<td>75.2</td>
</tr>
<tr>
<td>sysPass</td>
<td></td>
<td>1.3</td>
<td>50.3</td>
</tr>
<tr>
<td>Teampass</td>
<td></td>
<td>6.8</td>
<td>89.0</td>
</tr>
<tr>
<td>Websites</td>
<td></td>
<td>51.4</td>
<td>47.9</td>
</tr>
<tr>
<td>WordPress</td>
<td></td>
<td>46.0</td>
<td>49.9</td>
</tr>
<tr>
<td>Ghost</td>
<td></td>
<td>2.7</td>
<td>51.4</td>
</tr>
<tr>
<td>Cockpit</td>
<td></td>
<td>1.3</td>
<td>33.2</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>4.0</td>
<td>89.0</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Accessible from Internet</th>
<th>yes</th>
<th>via VPN</th>
<th>no</th>
<th>Other / unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smart Home</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home Assistant</td>
<td>21.5</td>
<td>25.3</td>
<td>0.0</td>
<td>59.4</td>
</tr>
<tr>
<td>Node RED</td>
<td>2.6</td>
<td>25.8</td>
<td>0.0</td>
<td>24.5</td>
</tr>
<tr>
<td>WebThings Gateway</td>
<td>1.4</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
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 self-hosters in the same use cases as self-hosters and non-self-hosters. Interestingly, Table 3 shows that across all use cases, self-hosters seem to use more tools in general. This means that they not only use more self-hostable 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>
<thead>
<tr>
<th>Table 3: User share per pre-defined tool (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tool</strong></td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Zoom</td>
</tr>
<tr>
<td>Discord</td>
</tr>
<tr>
<td>Microsoft Teams</td>
</tr>
<tr>
<td>Whatsapp</td>
</tr>
<tr>
<td>Telegram</td>
</tr>
<tr>
<td>Signal</td>
</tr>
<tr>
<td>Mumble</td>
</tr>
<tr>
<td>Teamspeak</td>
</tr>
<tr>
<td>Jitsi Meet</td>
</tr>
<tr>
<td>Rocket.Chat</td>
</tr>
<tr>
<td>Mattermost</td>
</tr>
<tr>
<td>Google Drive</td>
</tr>
<tr>
<td>Dropbox</td>
</tr>
<tr>
<td>Microsoft OneDrive</td>
</tr>
<tr>
<td>iCloud</td>
</tr>
<tr>
<td>MEGA</td>
</tr>
<tr>
<td>Box</td>
</tr>
<tr>
<td>Nextcloud</td>
</tr>
<tr>
<td>ownCloud</td>
</tr>
<tr>
<td>SparkleShare</td>
</tr>
<tr>
<td>Seafile</td>
</tr>
<tr>
<td>Syncthing</td>
</tr>
<tr>
<td>Amazon Alexa</td>
</tr>
<tr>
<td>Google Home</td>
</tr>
<tr>
<td>SmartThings</td>
</tr>
<tr>
<td>Home Assistant</td>
</tr>
<tr>
<td>Apple HomeKit</td>
</tr>
<tr>
<td>Node RED</td>
</tr>
<tr>
<td>WebThings Gateway</td>
</tr>
<tr>
<td>Bosch Smart Home</td>
</tr>
<tr>
<td>Domotzic</td>
</tr>
<tr>
<td>Gladys</td>
</tr>
<tr>
<td>Vivint Home</td>
</tr>
<tr>
<td>iCloud Keychain</td>
</tr>
<tr>
<td>LastPass</td>
</tr>
<tr>
<td>iPassword</td>
</tr>
<tr>
<td>Vault-Bitwarden</td>
</tr>
<tr>
<td>Roboform</td>
</tr>
<tr>
<td>Keeper</td>
</tr>
<tr>
<td>Dashlane</td>
</tr>
<tr>
<td>Padloc</td>
</tr>
<tr>
<td>Passbolt</td>
</tr>
<tr>
<td>Teampass</td>
</tr>
<tr>
<td>sysPass</td>
</tr>
<tr>
<td>WordPress</td>
</tr>
<tr>
<td>Blogger</td>
</tr>
<tr>
<td>Wix</td>
</tr>
<tr>
<td>Squarespace</td>
</tr>
<tr>
<td>Weebly</td>
</tr>
<tr>
<td>Ghost</td>
</tr>
<tr>
<td>Webflow</td>
</tr>
<tr>
<td>Strapi</td>
</tr>
<tr>
<td>Cockpit</td>
</tr>
<tr>
<td>Jimo</td>
</tr>
<tr>
<td>Wagtail</td>
</tr>
</tbody>
</table>

Bolded 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

USENIX Association
33rd USENIX Security Symposium 6063
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] and 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

<table>
<thead>
<tr>
<th>Operational classification</th>
<th>Non-Self-Hoster</th>
<th>Self-Hoster</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Self-Hoster</td>
<td>355 (60.3%)</td>
<td>122 (20.7%)</td>
<td>477 (81.0%)</td>
</tr>
<tr>
<td>Self-Hoster</td>
<td>32 (5.4%)</td>
<td>80 (13.6%)</td>
<td>112 (19.0%)</td>
</tr>
<tr>
<td>Overall</td>
<td>387 (65.7%)</td>
<td>202 (34.3%)</td>
<td>589 (100.0%)</td>
</tr>
</tbody>
</table>

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 \( w_h = \frac{N_h^{(1)}}{N_h^{(2)}} \) for each element in the \( h \)-th group defined by sex, age, ethnicity, and self-hoster status. \( N_h^{(1)} \) 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 (\( \alpha \)) for all scales requiring simple mean scores [1, 2, 12, 21, 24, 38, 68] and McDonald’s omega (\( \omega \)) 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
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 hosts, 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 Word-press 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


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


### 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>
<thead>
<tr>
<th>Sex × Age</th>
<th>Asian</th>
<th>Black</th>
<th>Mixed</th>
<th>Other</th>
<th>White</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 – 28</td>
<td>0.7</td>
<td>1.1</td>
<td>0.0</td>
<td>0.0</td>
<td>2.1</td>
<td>3.9</td>
</tr>
<tr>
<td>28 – 38</td>
<td>0.0</td>
<td>2.9</td>
<td>0.0</td>
<td>0.0</td>
<td>1.8</td>
<td>4.7</td>
</tr>
<tr>
<td>38 – 48</td>
<td>0.7</td>
<td>1.1</td>
<td>0.0</td>
<td>0.0</td>
<td>4.9</td>
<td>6.7</td>
</tr>
<tr>
<td>48 – 58</td>
<td>0.0</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
<td>4.4</td>
<td>5.0</td>
</tr>
<tr>
<td>58 – 150</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td>Overall</td>
<td>1.3</td>
<td>5.9</td>
<td>0.0</td>
<td>0.0</td>
<td>17.4</td>
<td>24.6</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 – 28</td>
<td>0.2</td>
<td>0.7</td>
<td>0.7</td>
<td>0.2</td>
<td>9.6</td>
<td>11.2</td>
</tr>
<tr>
<td>28 – 38</td>
<td>1.1</td>
<td>2.3</td>
<td>0.0</td>
<td>1.8</td>
<td>9.8</td>
<td>15.0</td>
</tr>
<tr>
<td>38 – 48</td>
<td>0.7</td>
<td>1.0</td>
<td>0.3</td>
<td>1.6</td>
<td>13.8</td>
<td>17.4</td>
</tr>
<tr>
<td>48 – 58</td>
<td>0.7</td>
<td>1.8</td>
<td>0.0</td>
<td>1.0</td>
<td>15.4</td>
<td>18.9</td>
</tr>
<tr>
<td>58 – 150</td>
<td>0.0</td>
<td>2.0</td>
<td>0.0</td>
<td>0.7</td>
<td>10.4</td>
<td>13.0</td>
</tr>
<tr>
<td>Overall</td>
<td>2.6</td>
<td>7.6</td>
<td>1.0</td>
<td>5.2</td>
<td>59.0</td>
<td>75.4</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 – 28</td>
<td>0.8</td>
<td>1.8</td>
<td>0.7</td>
<td>0.2</td>
<td>11.7</td>
<td>15.1</td>
</tr>
<tr>
<td>28 – 38</td>
<td>1.1</td>
<td>5.2</td>
<td>0.0</td>
<td>1.8</td>
<td>11.5</td>
<td>19.7</td>
</tr>
<tr>
<td>38 – 48</td>
<td>1.3</td>
<td>2.1</td>
<td>0.3</td>
<td>1.6</td>
<td>18.7</td>
<td>24.1</td>
</tr>
<tr>
<td>48 – 58</td>
<td>0.7</td>
<td>2.4</td>
<td>0.0</td>
<td>1.0</td>
<td>19.8</td>
<td>23.9</td>
</tr>
<tr>
<td>58 – 150</td>
<td>0.0</td>
<td>2.0</td>
<td>0.0</td>
<td>0.7</td>
<td>14.6</td>
<td>17.2</td>
</tr>
<tr>
<td>Overall</td>
<td>3.9</td>
<td>13.5</td>
<td>1.0</td>
<td>5.2</td>
<td>76.4</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Table 7: Share of sex × age × ethnicity groups in the population and our survey (in %)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 – 28</td>
<td>0.5</td>
<td>1.3</td>
<td>0.3</td>
<td>0.9</td>
<td>6.0</td>
<td>6.1</td>
<td>9.0</td>
<td>9.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28 – 38</td>
<td>0.6</td>
<td>1.2</td>
<td>0.3</td>
<td>0.8</td>
<td>5.8</td>
<td>5.8</td>
<td>8.6</td>
<td>8.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38 – 48</td>
<td>0.5</td>
<td>1.1</td>
<td>0.2</td>
<td>0.7</td>
<td>6.6</td>
<td>6.6</td>
<td>9.1</td>
<td>9.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48 – 58</td>
<td>0.5</td>
<td>1.2</td>
<td>0.1</td>
<td>0.5</td>
<td>7.4</td>
<td>7.3</td>
<td>9.7</td>
<td>9.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58+</td>
<td>0.6</td>
<td>1.5</td>
<td>0.2</td>
<td>0.3</td>
<td>12.4</td>
<td>12.5</td>
<td>15.0</td>
<td>15.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>2.7</td>
<td>6.3</td>
<td>1.1</td>
<td>3.1</td>
<td>38.2</td>
<td>38.3</td>
<td>51.4</td>
<td>51.5</td>
<td>48.6</td>
<td>48.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>48.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 – 28</td>
<td>0.5</td>
<td>1.3</td>
<td>0.3</td>
<td>1.0</td>
<td>6.4</td>
<td>6.3</td>
<td>9.5</td>
<td>9.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28 – 38</td>
<td>0.5</td>
<td>1.1</td>
<td>0.2</td>
<td>0.9</td>
<td>5.8</td>
<td>5.9</td>
<td>8.6</td>
<td>8.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38 – 48</td>
<td>0.5</td>
<td>1.1</td>
<td>0.2</td>
<td>0.7</td>
<td>6.6</td>
<td>6.6</td>
<td>9.0</td>
<td>9.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48 – 58</td>
<td>0.4</td>
<td>1.0</td>
<td>0.1</td>
<td>0.5</td>
<td>7.0</td>
<td>7.1</td>
<td>9.0</td>
<td>9.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58+</td>
<td>0.5</td>
<td>1.1</td>
<td>0.1</td>
<td>0.3</td>
<td>10.4</td>
<td>10.4</td>
<td>12.4</td>
<td>12.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>2.3</td>
<td>5.6</td>
<td>1.0</td>
<td>3.3</td>
<td>36.3</td>
<td>36.4</td>
<td>48.6</td>
<td>48.5</td>
<td>74.6</td>
<td>74.7</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Sex × Age</th>
<th>Asian</th>
<th>Black</th>
<th>Mixed</th>
<th>Other</th>
<th>White</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 – 28</td>
<td>0.0</td>
<td>5.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.2</td>
<td>2.2</td>
</tr>
<tr>
<td>28 – 38</td>
<td>0.0</td>
<td>12.5</td>
<td>0.0</td>
<td>0.0</td>
<td>3.4</td>
<td>3.9</td>
</tr>
<tr>
<td>38 – 48</td>
<td>12.5</td>
<td>± 24.5</td>
<td>0.0</td>
<td>0.0</td>
<td>6.1</td>
<td>5.1</td>
</tr>
<tr>
<td>48 – 58</td>
<td>0.0</td>
<td>14.3</td>
<td>± 28.0</td>
<td>0.0</td>
<td>14.3</td>
<td>14.3</td>
</tr>
<tr>
<td>58+</td>
<td>0.0</td>
<td>2.6</td>
<td>± 5.0</td>
<td>0.0</td>
<td>2.9</td>
<td>2.9</td>
</tr>
<tr>
<td>Overall</td>
<td>2.6</td>
<td>0.0</td>
<td>± 5.0</td>
<td>0.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

| male      |       |       |       |       |       |         |
| 18 – 28   | 14.3  | ± 28.0| 5.0   | ± 9.8 | 0.0   | 17.2    |
| 28 – 38   | 37.5  | ± 35.9| 41.2  | ± 24.1| 23.1  | 11.9    |
| 38 – 48   | 14.3  | ± 28.0| 0.0   | ± 9.8 | 11.1  | 12.6    |
| 48 – 58   | 0.0   | 26.7  | ± 23.2| 0.0   | 22.1  | 21.1    |
| 58+       | 0.0   | 20.0  | ± 21.0| 0.0   | 8.4   | 9.3     |
| Overall   | 14.3  | ± 11.5| 18.0  | ± 7.9 | 12.3  | 14.3    |

| Overall   |       |       |       |       |       |         |
| 18 – 28   | 7.1   | ± 14.0| 5.0   | ± 6.9 | 19.7  | 9.7     |
| 28 – 38   | 17.5  | ± 16.8| 26.9  | ± 14.7| 12.2  | 7.6     |
| 38 – 48   | 13.3  | ± 18.5| 0.0   | ± 9.8 | 5.5   | 9.3     |
| 48 – 58   | 0.0   | 12.5  | ± 10.8| 0.0   | 15.4  | 12.2    |
| 58+       | 0.0   | 8.0   | ± 8.4 | 0.0   | 9.2   | 4.7     |
| Overall   | 8.0   | ± 6.0 | 10.0  | ± 4.1 | 5.9   | 8.3     |

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