Fix it - If you Can! Towards Understanding the Impact of Tool Support and Domain Owners' Reactions to SSHFP Misconfigurations

1st Anne Hennig
Karlsruhe Institute of Technology
Karlsruhe, Germany
anne.hennig@kit.edu

2nd Sebastian Neef
TU Berlin
Berlin, Germany
neef@sect.tu-berlin.de

3rd Peter Mayer
South Denmark University
Odense, Denmark
mayer@imada.sdu.dk

***Abstract*—Misconfigured SSHFP records might lead to SSH users not carefully verifying host key fingerprints, making SSH connections vulnerable to Man-in-the-Middle attacks. To warn domain owners about SSHFP misconfigurations and the potential security implications, we conducted a 2×3 randomized controlled notification experiment. We sent notifications to $n = 518$ domain owners with misconfigured SSHFP records. Following up on contradictory results from related work, we investigated the effects of tool support. While we see that the sender of the notification itself has no effect, our results suggest that tool support might increase remediation when the sender of the notification is different than the institution providing the tool. Furthermore, we analyzed domain owners' responses to our notification to identify reasons for (non-) remediation. While only 27% remediated the misconfiguration after our notification, we identified valuable explanations for individual remediation behavior in the responses we received ($n = 52$), supporting the argument that remediation rate should not be considered a success measure for a notification campaign but instead individual challenges faced by domain owners should be taken into account.**

***Index Terms*—DNSSEC, SSHFP, SSH, misconfiguration, notification experiment, vulnerability notification**

1. Introduction

The Secure Shell Protocol (SSH) was designed to allow access to remote systems over insecure network connections [1]–[7]. SSH relies on TOFU (Trust On First Use) and PKC (Public-Key Cryptography) to authenticate an SSH server. Upon the first connection to an SSH server, which usually happens via a SSH client, SSH users should validate the fingerprint of the server's public key through an out-of-band channel, e.g., by obtaining it from the server administrator. However, SSH users tend to not properly follow this crucial security-relevant step, called Host Key Verification (HKV), and just accept the displayed fingerprint, opening attackers avenues for compromising the connection or system [6], [8]. Under certain circumstances, a network-based attacker could steal password-based authentication credentials [9], [10], or deanonymize users through their public keys [11].

One solution to facilitate HKV for SSH users was standardized in RFC 4255 [6]. It introduced a new Domain Name System (DNS) resource record type named "SSHFP" (Secure Shell Fingerprint) to store host key fingerprints and allows SSH clients to query the server's Fully Qualified Domain Name (FQDN) for these records to verify the authenticity of the server's host keys automatically. However, this requires the SSHFP records to be transmitted tamper-resistant, e.g., by activating DNSSEC for the domain. Related work (e.g., [12], [13]) has shown that these SSHFP records are not always correctly configured, which makes SSH clients fall back to manual HKV.

In order to make domain owners, i.e., the individuals who registered a domain or are responsible for its content, aware of the potential security risks due to misconfigured SSHFP records, the main goal of our study is to notify affected domain owners. However, this raises the question of how to reach them best, and how to get them to act upon a notification. Given that we report a misconfiguration rather than a critical security issue, it seems promising to notify domain owners individually and in a targeted manner. Many notification experiments have been conducted in recent years to find the most effective senders (i.e., individuals such as security researchers who send a vulnerability notification), notification channels, notification texts, or additional support in the form of, e.g., a more detailed report on a website or a self-test-tool [14]–[28] for notifying individuals about security issues or misconfigurations in their systems. To gain further insights into factors that help design misconfiguration notifications, we conducted a mixed-methods study consisting of a notification experiment and thematic analysis of the email responses we received.

Our first goal was to find out which of the previous results related to sender and additional support in the form of a self-test tool (i.e., tool support) can be replicated when investigated in the same study setting, eliminating the need to compare results across different studies. Secondly, this study design allowed us to be the first to investigate possible interaction effects between the tool provider, i.e., the institution that provided the self-test tool, and the sender of the notification, in our study researchers from two German universities. Thirdly, we examined the responses to our notifications in more detail, as [18], [25] suggested

that remediation rates might not be sufficient to explain remediation behavior.

Specifically, our work consists of four research questions, which we motivate in detail in Section 3:

RQ 1 [Tool Support] *Does an interactive self-test tool increase remediation compared to textual information?*

RQ 2 [Sender x Tool] *Does an interaction effect exist between the provider of the tool and the sender of the notification?*

RQ 3 [Response] *What are the reasons for domain owners' non-remediation of the misconfiguration?*

RQ 4 [Reaction] *How do domain owners react to notifications about misconfigurations?*

2. Background

Secure Shell and Host Key Verification. The Secure Shell Protocol (SSH) was designed to allow access to remote systems over insecure network connections [1]–[7]. The protocol was designed to rely on Trust On First Use (TOFU) for checking the authenticity of the remote server upon connecting. That is, during the first connection to a new server, the user is asked to manually verify the server's host key fingerprint, which should be obtained from or provided by the administrator through an out-of-band channel. This "Host Key Verification" (HKV) step is used to ensure the users are connecting to the correct server. Afterwards, SSH clients such as OpenSSH will store the verified host key locally and use it for subsequent connection attempts. Automated HKV without user interaction can be achieved with DNS-based HKV, as described below in more detail.

Domain Name System and DNSSEC. The Domain Name System (DNS) was initially standardized in RFC 882 [29] and has become an indispensable part of today's internet usage, as it allows to resolve rememberable domain names to IP addresses. Furthermore, different DNS resource record types were standardized in the past to extend DNS use cases. For example, `A` and `AAAA` records are for IP addresses, while `MX` records are used to identify mail servers. However, as a plaintext protocol, DNS is prone to network-based attacks, such as the manipulation of DNS records in transit, i.e., by Man-in-the-Middle (MitM) attacks. The DNS Security Extensions (DNSSEC) standard (RFC 9384 [30]) brings authenticated records to DNS. While the usage and adoption of DNSSEC comes with its own challenges (e.g., [31], [32]), it allows DNS records to be relied upon for security-related processes, such as DNS-based HKV.

DNS-based Host Key Verification. RFC4255 [6] introduced the *SSHFP* DNS record type, which allows to place a SSH server's host key fingerprints in the DNS to enable automated HKV, thus, removing the risk of verification mistakes due to human error [6], [8].

There are two requirements for DNS-based HKV to work securely, as depicted in Figure 1:

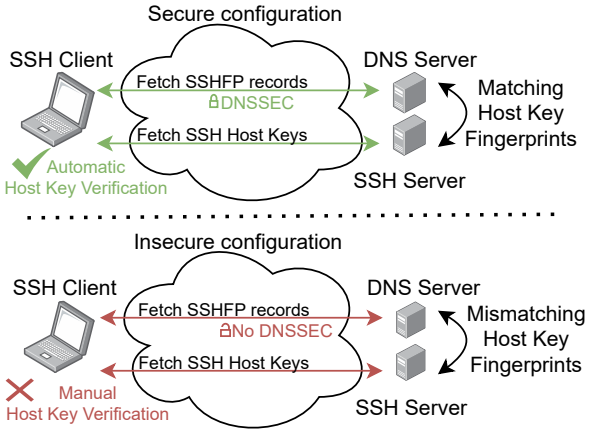


Figure 1: Secure vs. insecure configuration of SSHFP records leading to manual host key verification.

(1) The host key fingerprint(s) in the SSHFP record(s) must match with the server's actual host keys.
(2) The SSHFP records are always transferred and validated using DNSSEC. If local validation is not possible, the records need to be securely transmitted between the client and the validating resolver.

Thus, in a secure configuration, the SSH client can fetch the DNSSEC-secured SSHFP records containing the correct host key fingerprints and then successfully validate these against the SSH server's host keys. If the comparison succeeds, the authenticity of the host is established without user interaction, even upon the first connection to the server (see *secure configuration* in Figure 1).

Misconfigured Host Key Verification. On the other hand, there are two situations in which this automatic process can fail, as depicted in the *insecure configuration* scenario, whose remediation can vary in complexity:

(1) Mismatching host key fingerprints: This issue can be remediated by an administrator by updating the respective SSHFP DNS records with the correct host key fingerprints.
(2) Missing DNSSEC validation: DNSSEC needs to be enabled for the domain by an administrator, which may require more configuration efforts and coordination with upstream DNS providers to avoid wrong or insecure configuration [32].

In both cases, a user's SSH client and system need to be instructed to perform SSH-based HKV, including correct DNSSEC validation, which might impose another challenge from an administrator's perspective. If automated HKV is not possible, the OpenSSH client falls back to manual verification again. If a SSH user does not correctly validate the host keys to ensure they are connecting to the intended SSH server and automatic HKV is not possible, a suitable positioned, network-based attacker might be able to obtain login credentials or learn information about their victim depending on the configuration [9], [11].

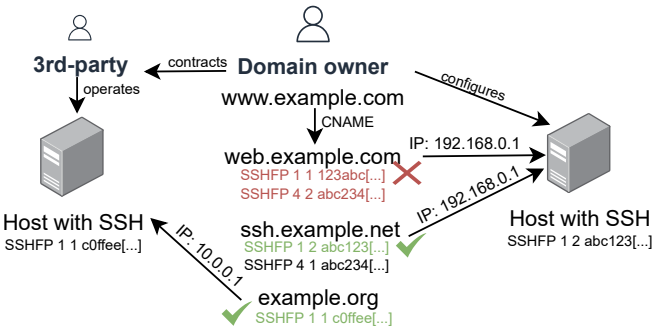


Figure 2: Illustration of the relationship between a domain owner, their domains, third-parties, SSHFP records and hosts.

While this attack scenario appears to be feasible in theory, in practice it will require several conditions (e.g., password-based authentication, manual HKV, MitM-capable attacker, first connection to a server) to be met, likely limiting its practicability outside lab environments. However, related work suggests that some of the required conditions are imaginable. Andrews et al. [33] have shown that, over 65% of the internet-reachable SSH servers had password authentication enabled. Neef et al. [13] reported a significant number of mismatched SSHFP records and a frequent lack of DNSSEC protection for these records. Furthermore, Thankappan et al. [34] reviewed MitM-attacks and concluded that network packet manipulation is still feasible and not fully mitigated in WiFi protocols.

Therefore, the notification campaign described in this work aims to make the respective domain owners aware of the potential risk of incorrectly configured SSHFP records, which should prompt them to review and update their configuration. Choosing SSHFP misconfigurations for this study appeared sensible as little prior research on this exact record type and its misconfigurations exists, warranting a closer examination to gain a better understanding of it. Further, SSH's prominent role in server administration and the discussed impact reinforce the importance of raising awareness of incorrectly configured SSHFP records.

Figure 2 illustrates the relationship between the different entities used in the remainder of the paper. A domain owner can have multiple second-level domains (i.e. `example.com`) having SSHFP records set for the second-level or any sub-domain individually. Sometimes, the servers or domains are not under the domain owner's direct control, as they might be hosted on shared hosting services, meaning that the domain owners have to contact a third-party (e.g., their hosting provider) to get the records changed. The latter also applies to CNAME entries, which point to domains not under the domain owner's control. In any case, the fingerprints should match the SSH host-keys on the server pointed to by the domain's IP address (A record).

3. Related work

3.1. SSHFP Studies

While there has been active research about SSH (e.g., [35]–[41]), the DNS-based HKV using SSHFP records (RFC4255 [6]) has only seen limited coverage. For example, Gasser et al. [12] analyzed SSH servers and SSHFP records in 2014, and Neef et al. [13] followed up with a large-scale analysis of these records' configuration in 2022. Both studies reported a low adoption rate of SSHFP records. Additionally, Neef et al. discovered up to 76% misconfigured domains, whose SSHFP records do not fully match the system's actual host keys, and DNSSEC support for only up to 39% of the analyzed domains. To the best of our knowledge, neither Gasser et al. nor Neef et al. notified the affected domain owners, and we are not aware of a notification study that raised awareness among domain owners about misconfigurations of the SSHFP records and missing DNSSEC.

3.2. Notification Studies

Previous work has shown that reaching out to affected individuals[1] is not an easy task. Several conditions have already been tested to find the most effective ways - comparing different senders or recipients (e.g., [16], [17], [19]–[21], [23]–[25], [28]), comparing message texts and length (e.g., [23], [24], [42]), and comparing different notification channels (e.g. [16], [18]–[20], [26], [27]).

Sender and Recipient. Concerning the recipient, direct (e.g., directly via email [19], [20], [23], or via customer support accounts [16], [24], [26], [27]) and indirect (e.g., hosting provider [21], [23], [25], [28], or CERTs [15], [17], [21]) contacts were tested. Direct approaches with manual or automatic retrieval of contact information (e.g., from the imprint of a website) proved most effective [18]–[20], [42]. Thus, for our study design, we decided to manually retrieve contact information from the websites and contact domain owners directly.

With respect to the sender, individuals who did not disclose themselves as researchers [19], individual researchers [20], [21], [28], university groups (e.g. university computer science group [19], [28], or university law group [19], [25]), or external senders (e.g. anti-malware organizations [28], or generic senders [20], [24]), were tested. Besides the university law group [19] no sender proved to be more effective than the other senders compared in the respective studies. Thus, for our study design, we decided to sent notifications from two individual researchers affiliated with two major German universities (Technical University of Berlin (U1) and Karlsruhe Institute of Technology (U2)).

1. Depending on the issue, these were, e.g., system operators, website owner, domain owner or system administrators. For our study, these were domain owners.

Tool Support. For notification content and providing verification possibilities, tool support[2] has been proposed in previous studies [16]–[19], [25], [43]. Tool support has not been proven effective in increasing remediation rates compared to a not notified control group (i.e., all treatment groups received the link to the tool) [19], nor proven effective with respect to remediation rates compared to text-only notifications when notifying nameserver operators, domain owners, and Internet Service Providers (ISP) [25]. However, a significant amount of notifications to the domain owners in Cetin et al.'s [25] study bounced (slightly more in the demonstrator group), and the authors mentioned that the demonstrator website seemed to have some usability issues. Furthermore, providing a self-test tool as recipients appreciate verification, increased remediation rates (where used) and led to a more permanent remediation, and is, therefore, recommended by previous work [18], [19], [25], [43]. Understanding the effects of tool support can help clarify why self-test tools are perceived as useful by affected individuals and researchers, even when they have yet to show significant effects on remediation. By comparing a tool to a text-only notification in a different context than used by Cetin et al. [25] we contribute valuable insights on when a self-test tool is effective (**RQ 1**).

Maass et al. [19] analyzed the effectiveness of a self-test tool that was hosted on the servers of the University of Bamberg, while the senders of the notifications were affiliated with the University of Darmstadt. However, their study had no baseline group with respect to tool support, i.e., all notified users received the link to the tool. The authors found that although the self-test tool did not increase remediation rates significantly, users nonetheless found it helpful. This discrepancy raises the question of underlying factors that might influence an affected individual's decision to use such a self-test tool. Specifically, it suggests that interaction effects might exist between the institution the sender of the notification is affiliated with and the institution hosting the tool. It might be that, e.g., a tool hosted at another institution might provide further credibility. On the other hand, a link to an external tool provider might look suspicious and is, thus, not used. The fact that it has not yet been researched whether this separation between tool provider and sender of the notification was beneficial for the notification campaign motivated us to investigate possible interaction effect between tool provider and sender of the notification (**RQ 2**).

Responses. Another topic that came up in previous studies was how meaningful the interpretation of remediation rates is without considering the affected individual's perspective. While not analyzing responses in depth, Cetin et al. [25] discussed that just because an issue is not remediated does not imply that affected individuals are not aware or careless. Stöver et al. [44] thematically analyzed the responses from Maass et al. [19] and found that website owners face multiple challenges that can be, for example, lack of knowledge, but also lack of resources or complex organizational coordination that slow down or hinder remediation processes. Understanding the reasons for non-remediation and examining the challenges for remediation will help future work to tailor notifications to affected individuals in the best possible way. Thus, we explored reasons for domain owners not to remediate the misconfiguration (**RQ 3**).

Furthermore, it has been shown that responses to security issues differ from those to privacy issues [42]. As described by Maass et al. [19] reactions to notifications on privacy related misconfigurations can range from requesting help to complaints or just gratitude. By analyzing the responses we received in the context of security related misconfigurations, we contribute to the current findings in that we provide further insights on the motivations and reactions of different affected individuals whom researchers are addressing in the context of notification campaigns (**RQ 4**).

4. Methodology

4.1. Sampling

Domain Sample. For our study, we focused on domains belonging to the *.de* ccTLD (country code top-level-domain). At the time of writing, the German TLD operator DENIC reported almost 17.7 million registered domains [45]. While several TLD providers make their domain data freely accessible [46], this is not the case for the DENIC. Therefore, we gathered domains from publicly available sources as used by other measurement studies (e.g., [13], [47], [48], [48]):

- OpenINTEL's [46] popular domain lists, such as the Alexa Top 1M, Cloudflare Radar Top 1M, Tranco 1M and Cisco Umbrella Top 1M, resulting in 41,206 unique *.de* domains with 101,322 subdomains;
- CRT.sh[3], resulting in 260,158 unique subdomains belonging to 117,009 *.de* domains;
- CertStream[4], resulting in over 2.2M unique *.de* domains with 4.5M associated subdomains;
- Censys [51], resulting in almost 10M unique *.de* domains and over 31M subdomains;

In total, we collected over 12.3M domains and 36M subdomains from these sources, of which 9.98M domains and 31.1M subdomains were unique after removing duplicates. The domains in this dataset were not yet validated and, thus, included expired domains or domains with wildcards (e.g., `*.example.com`), which were unsuitable for further analysis. Consequently, we eliminated wildcards and

2. Note, that while we call it "tool support" we do not refer to a software product that the domain owners need to install. Instead, we built a demonstrator website where the domain owners can verify the problem with a self-service web application and test whether they have remediated the issue (see Section 4.3 for more information).

3. An online service that allows querying of certificate transparency logs [49].

4. Another service that gathers information about issued certificates in real-time [50].

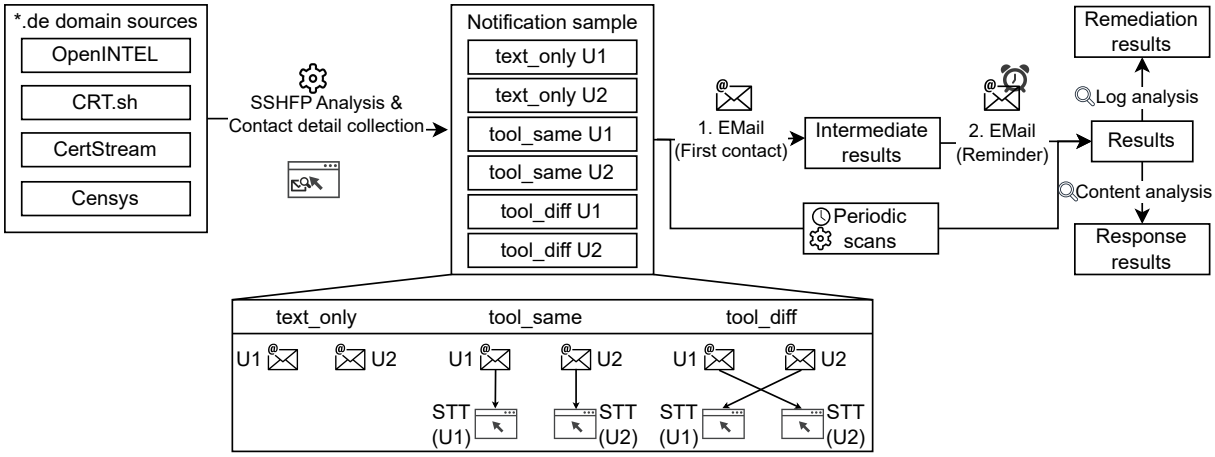


Figure 3: The methodology of our notification experiment, ranging from the domain sampling to analyzing the results

validated the existence of each domain, resulting in a dataset of resolvable and active[5] German domains.

The final domain dataset comprised 7.97M domains and 26.5M subdomains, representing around half of the registered German domains (7.97M / 17.7M = 45%) according to [45]. We make these datasets and related code available for future work in our repository [52].

Notification sample. We used Neef et al.'s analysis prototype [13] with slight modifications to obtain our notification sample. The modifications include longer DNS query timeouts ($t = 10s$) and a slower DNS query rate ($r = 10q/s$) to improve the DNS scanning reliability as well as an additional analysis step to output domains and their respective misconfiguration issue in a separate file.[6] These issues include *mismatching fingerprints*, where the fingerprints in the SSHFP record do not match the server's actual host keys, and *missing DNSSEC*, where the domain does not have DNSSEC configured. After scanning all 26.5M subdomains, we found a slightly higher, but still low adoption rate of SSHFP records in our German sample compared to Neef et al. [13] with only 10,462 (0.04%) subdomains belonging to 1,289 domains setting this record type (compared to 0.011% for the Tranco 1M and 0.013% for the CT datasets in [13]). The discovery of an SSH server on the default port (22/tcp) for the host key comparison succeeded for 3,267 subdomains. Over 50% (1,638) of these subdomains, which belong to 498 unique domains, have an insecure configuration, due to *mismatching fingerprints* (437 subdomains, 168 domains) or *missing DNSSEC* (1201 subdomains, 330 domains). Interestingly, we not only see a slightly higher SSHFP adoption rate in our German sample, but also less misconfiguration compared to the international sample in Neef et al. [13]). Figure 7 in Appendix E provides a visualization of the statistics.

5. We queried the `.de` TLD's DNS servers for a first-level domain's authoritative nameservers. If we received a non-empty set, we assumed the domain to be valid and active.

6. A repository with code and data is available at [52].

For our study, we decided to notify domain owners about the misconfiguration, as we assumed they would have the authority to change SSHFP records for the misconfigured (sub-)domains. In order to identify the contact information in the form of email addresses of the affected domain owners, we first mapped subdomains to their respective domains and grouped them, e.g., when different subdomains or even domains displayed the same content and, thus, obviously belonged to the same domain owner. One challenge that we encountered was when different subdomains of a domain hosted individual websites with distinct contact information, as was the case for universities, where some departments had their own websites with their own imprints. In such cases, we treated the subdomains as separate cases since the notification would be addressed to – and likely be processed by – different individuals.

We deliberately chose email as the only notification channel for our experiment. The main reason for this choice is that none of the previous studies have proposed a feasible alternative for large-scale notification campaigns [18]–[20]. All studies criticize the fact that letters [18], [19] or phone calls [20] are too inefficient and that the enormous amount of time and money spent on alternative contact channels is disproportionately high compared to the (small) increase in remediation. We also deliberately decided against social media [20] as a notification channel as this seems not to be an appropriate way for researchers to contact domain owners about a misconfiguration.

Contact information, i.e. the email addresses and names of domain owners, were collected by manually visiting the imprint or the contact information page from a domain's website, as in [43], [53]. We collected the email addresses given for the persons being responsible for the website (i.e., usually the person being named in the first paragraph as, e.g., "service provider", "responsible person", or "operator"). If the email contact of a webmaster was given, we used theirs. If we were unable to retrieve this

information or there was no website present[7] we used the standard address hostmaster@<domain> (for example, hostmaster@example.com) for DNS related requests as specified in RFC 2142 [54].

We also checked all misconfigured domains for security-related contact information in *security.txt* [55]. As its adoption is generally limited especially for less popular websites [56], this approach only revealed useful contacts for nine domains in our sample (9/498=1.8%)[8]. To avoid adding further variables to our study design, we only used the email addresses we manually retrieved by visiting the imprint of each website or generated according to RFC 2142 to keep the study's design similar to related work (e.g., [25], [53]). In total, we identified email addresses for 518 cases (304 manually collected, 214 generic), which formed our notification sample and which we notified about the misconfiguration.

4.2. Study Design Notification Experiment

To evaluate effects and interaction effects of sender and tool support, we designed a 2×3 randomized controlled notification experiment. We defined two individual researchers from the computer science departments of two different universities in Germany as senders. Both senders used their institutional email address using their names and their institution's abbreviation as sender identity ([firstname.]lastname@<domain>[9]). Since we did not have direct control over outgoing mail servers' configuration, we contacted our universities' IT departments. We were assured that both mail servers are configured similarly and to the state-of-art best practices (i.e., SPF, DKIM, DMARC) to maximize email acceptance and minimize the risk of our notifications being marked as SPAM. We send out the emails at a slow rate and as U1's and U2's email servers are supposed to handle large volumes of incoming and outgoing emails in short times, we did not expect our notifications to be blocked.

With respect to the content of the notification, we used best practices provided in [43], [57], e.g., using a personal salutation, clearly stating our motivation, providing detailed information, pointing out possible consequences if the misconfiguration is not remediated, and providing the senders' names, affiliations, and further contact information (see Section C.1, Figure 6 for the notification text).

To answer **RQ 1** and **RQ 2**, we designed three different notification types: text_only, tool_same, tool_diff (see Figure 3). The text_only condition contained only the text described above. The tool_same and tool_diff conditions additionally contained a link to the self-test tool (STT; see Section 4.3). In total, we defined six groups for our experiment (see "notification sample"-box in Figure 3):

- Two groups that received the email with the textual explanation but without a link to the STT (text_only U1, text_only U2).
- Two groups that received the email with the textual explanation including a link to the STT hosted by the sender's institution (tool_same U1, with link to STT hosted at U1; tool_same U2 with link to STT hosted at U2).
- And two groups which received the email with the textual explanation, including a link to the STT hosted by the other sender's institution (tool_diff U1, with link to STT hosted at U2; tool_diff U2, with link to STT hosted at U1).

The domains were evenly distributed randomly to one of the six groups (see Figure 3). Table 2 in Section C.2 contains the final numbers of domains in each group.

We used the MailMerge Add-on[10] in Thunderbird to prepare the emails for sending. We used the option to save the personalized email as draft, in order to check each email for correctness and to not trigger SPAM or other filters by sending them in bulk. Both authors sent out their share of 518 notifications each between 12.30 and 1 pm on March 6, 2024. Since the emails were sent out manually, this process took approximately 30 minutes per sender (on average $\leq$ 9 emails/minute). Three weeks later, on March 26, 2024, between 5.30 and 6 pm a reminder was sent using the same process to those domain owners who had not remediated the misconfiguration or had not answered our initial email yet (see Appendix C.1, Figure 6 for the reminder notification text). We monitored the domains for another two and a half weeks, until April 12, 2024 to detect delayed remediation (for more details on the monitoring, see Section 4.4).

4.3. Self-Test Tool (STT)

In order to provide affected domain owners with the ability to verify the problem we described, and check whether their changes to the configuration were successful, we implemented a web-based self-test tool (STT)[11], similar to [19], [25]. Our STT implements the same validation checks and uses the exact same code as our initial analysis script (see Section 4.1). To make the script's output more user-friendly, we displayed the results in a structured way, using different colors to indicate whether a configuration is correct or not, in addition to the scan results (see Figure 5 in Appendix B).

Two identical instances of the STT were made available at subdomains of the websites of U1 and U2, e.g. ssh-security-check.<domain>. When accessing the respective website, it first showed the FAQ page, where we provided answers to 16 questions addressing the

7. Note, that some (sub-)domains were not used as a webserver but, e.g., as a mailserver, so it was not unusual that a domain had no or an empty website.

8. We requested both, "/security.txt" and "/.well-known/security.txt" over "https://" in accordance with RFC 9116 [55]. Returned pages were checked for (email) contact details in "Contact:" as per RFC 9116. See [52] for more information.

9. U1 only uses the lastname, whereas U2 uses firstname.lastname

10. https://addons.thunderbird.net/thunderbird/addon/mail-merge/

11. The links will be added in the final version. The code is available in [52].

domain owners' potential concerns, describing our motivation, providing contact details, and recommending remediation steps. The first FAQ entry guided the domain owner to a subpage with the STT. To avoid malicious or unintended use, the STT could only be used if a valid token was entered[12]. We generated a random alphanumeric 8-character token for each notified domain owner in the `tool_same` and `tool_diff` groups. Each website owner could, thus, only check the (sub-)domains that belonged to the respective token. If the token was incorrect or did not match the (sub-)domain that was entered, no scan was performed by the STT. The website was made available in German and English to mitigate respective language barriers. We did neither publicly advertise the website nor the STT.

4.4. Data Analysis

Periodic Scans. To identify when an insecure configuration was remediated, we set up continuous monitoring throughout the evaluation period of 37 days ($\sim$ 5 weeks). Every six hours (at 00, 06, 12, and 18 o'clock), all (sub-)domains belonging to the notified individuals were checked automatically using the scanning tool described in Section 4.1. The resulting data was then used for the statistical analysis.

Statistical Analysis. For our study, we defined two independent variables (sender and tool support) with the sender having two nominal characteristics (U1 and U2), and tool support having three nominal characteristics (`text_only`, `tool_same`, `tool_diff`), which resulted in total in six characteristics when assessing interaction effects (see Figure 3): `text_only` U1, `text_only` U2, `tool_same` U1, `tool_same` U2, `tool_diff` U1, `tool_diff` U2. As described in Section 4.2, we controlled these variables by evenly distributing the domains to the treatment groups. Our dependent variable was remediation (i.e., the correction of the configuration), which we measured a continuous variable (time until remediation) and as dichotomous variable (status at the end of the study period).

We analyzed the differences between the treatment groups (**RQ 1**), with a single-factor ANOVA (days until remediation as dependent variable) and a χ^2-test (status as dependent variable). We used two-factor ANOVA (days until remediation as dependent variable) to identify interaction effects between the institution hosting the tool and the sender of the notification (**RQ 2**). Furthermore, we used survival analysis to analyze the remediation behavior over time similar to several previous studies in the context of notifications experiments [17], [19], [23], [24], [26], [28].

To determine the necessary sample size for our statistics, we used an a priori power analysis [13]. Since no effect sizes were reported in related work, we assumed a medium effect for all tests (f = .25, resp. w = 0.3). Assuming a medium effect size is common practice when calculating sample sizes. While assuming a medium effect, might leave studies with the risk of being underpowered, one could argue that assuming a low effect size without good reason could raise the question why an investigation, i.e., the notification in our case, would even be necessary if it is assumed to be only little effective.

The results of the a priori power analysis with G*Power showed that a sample size of 324 notifications is recommended for single-factor ANOVA with six groups. A sample size of 220 notifications is recommended for χ^2-test. For two-factor ANOVA that we used to determine interaction effects between the sender and the institution providing the STT we calculated a total sample size of 400 notifications. We could not calculate the sample size for the survival analysis with G*Power. Since survival analysis is based on Cox regression as a semi-parametric model, we determined the sample size based on logistic regression. Following Bujang et al. [58] we determined that we require $n = 100{+}50{\cdot}\#independent\ variables = 100{+}50{\cdot}6 = 400$ notifications. Our sample was supposed to exceed all these thresholds ($n = 518$ notifications sent). However, since 132 emails were not delivered and 36 notifications needed to be deleted, our analysis sample only included 350 notifications. All statistical tests were conducted using an alpha level of .05, and post-hoc Holm-Bonferroni correction was applied to counter alpha error cumulation where necessary.

Thematic Analysis of Responses. Finally, we used thematic content analysis to answer **RQ 3** and **RQ 4**. We received 52 responses from the 518 domain owners we notified. We excluded two answers, one who explicitly asked that their data should be excluded, and one who answered to both senders (we discuss this in Section 6.3). For the thematic analysis, we manually removed all personally identifiable information, and only analyzed the anonymized texts.

We used a codebook proposed by Stöver et al. [44]. While their codebook refers to "web operators", we could easily transfer the codes so that they fit our purpose, e.g., by slightly modifying the original wording (e.g., we exchanged "reasons for lack of AIP" with "reasons for lack of remediation"), and adding detailed explanations. We also added codes for sentiment analysis (i.e., *tone* and *content*) to compare the responses we received to other notification experiments (e.g., [19], [42]). We discussed our codes with the authors of [44] to make sure our adapted codebook is still comparable to the original one.

To assure inter-rater reliability (IRR), calculated using Cohen's Kappa, two coders manually coded 8% of the material together. They then met, discussed ambiguities, and refined the codebook. Again, 4% of the material was coded together. Cohen's Kappa was then calculated, reaching an IRR above 0.7 ($\kappa = 0.74$). This was deemed sufficient, so the rest of the material was coded manually by one of the coders (see Appendix F.2 for our codebook).

12. A similar approach was used by Cetin et al. [25].

13. Note that our sample size was naturally determined by the number of misconfigured domains we were able to detect and, in the end, even more restricted by our exclusion criteria. Nevertheless, we think an a priori power analysis was helpful to properly discuss possible limitations of our results.

4.5. Ethical Considerations

While designing our study, we carefully considered the important lessons learned from previous notification experiments. For example, we decided to use a misconfiguration as the basis for our notification, which does neither represent an immediate threat nor introduces a vulnerability, as used by other studies, e.g., [59]. Additionally, to the best of our knowledge, the reported misconfiguration does not entail any legal consequences or legal obligations for the domain owners so that we could sidestep potential drawbacks of fear, anxiety, anger, or mistrust among affected individuals, as reported by [60].

Reminder notifications could also increase mental overload, as discussed in [17]. However, since Maass et al. [19] could prove the effectiveness of reminder notifications, we decided to send one reminder notification, but only contacted those domain owners again who did not respond or remediate three weeks after our initial notification.

The notification included our contact details, and the information that the notification is part of a combined research project from U1 and U2. For those receiving the link to the STT, the corresponding website included an extensive FAQ answering potential questions the domain owners might have about our study [14]. Thus, no deception occurred, and we provided sufficient information that the email was sent as part of a research project. The domain owners could opt-out of any follow-up communication or data processing at any time if they answered one of our notification emails. The opt-out was requested and honored in one case, which is in contrast to the many positive responses as discussed in Section 5.2.

To the best of our knowledge, we followed the Guidelines for Safeguarding Good Research Practice and further guidelines issued by the ethics commission of U1 and U2 to ensure that our study does not pose any potential harm or risk to the domain owners. Note that according to U2, our study was exempted from ethics oversight, and for U1 our study design did not require IRB approval.

5. Results

5.1. Results of the Notification Experiment

Our notification sample contained in total 518 domain owners that we notified. For our analysis we needed to exclude 132 cases where the email notification was not delivered (118 `hostmaster@<domain>`, 14 manually collected), one domain where the domain owner asked to be excluded, and 35 domains that we learned during the notification process belong to the same domain owner, although having different contact information in the imprints. Thus, our analysis sample contained 350 notifications.

We created two datasets with two subsets each based on different interpretations of remediation: `DS-1` is based on the *domain-level*, where each domain was treated as one case if all subdomains showed the same remediation behavior. We excluded all domains where the remediation behavior differed for individual subdomains. This dataset is closest to our notification sample and reflects remediation without any interpretation, based only on the scan results. `DS-2` is based on *subdomain-level*, where we treated those subdomains as one case that showed the same remediation behavior. This means that for some domains we then had several cases. This dataset reflects the researcher's interpretation that domain owners made individual decisions on a subdomain basis, i.e., implies that our notification is effective even if the configuration has not changed for the whole domain but only for certain subdomains.

From both datasets we created two subsets:

- `DS-1-RT` ($n = 302$) and `DS-2-RT` ($n = 338$) to analyze remediation over time (***RT***). In this subset, domains with periodic scan errors[15] for more than four consecutive data points (i.e., when a whole day's data was missing during the 37 days observation period) were excluded to ensure a coherent survival analysis.
- `DS-1-RS` ($n = 318$) and `DS-2-RS` ($n = 348$) to analyze the final remediation status (***RS***) at the end of the study period. For this dataset, only domains with four consecutive errors in the last four data points (i.e., no data on the last day of the study period) were excluded, but not the ones having scan errors during the observation period.

Remediation Rates. The average remediation rate for our notification experiment was 26.0% for `DS-1` (76/302 = 25,2% remediated in `DS-1-RT`, 86/318 = 27.0% remediated in `DS-1-RS`), and 29.0% for `DS-2` (97/338 = 28,7% remediated in `DS-2-RT`, 101/348 = 29.0% remediated in `DS-2-RS`), meaning that at least 70% of the domains in our sample did not change their configuration based on our notification. The mean remediation time ranged between 96.95 scans (`tool_diff` U1, approximately 24 days) and 131.15 scans (`text_only` U2, approximately 33 days) for `DS-1`, and 96.76 scans (`tool_diff` U1, approximately 24 days) and 131.19 scans (`text_only` U2, approximately 33 days) for `DS-2`.

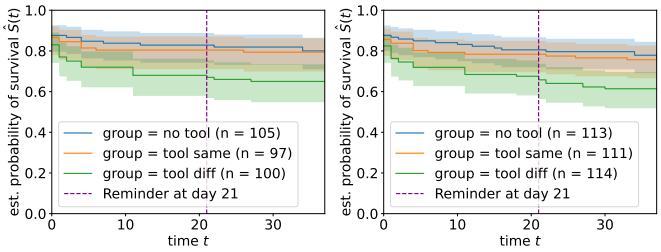
Not surprisingly, we found a significant difference between the notification types *mismatching fingerprints* and *missing DNSSEC* in remediation status ($\chi^2(1) = 19.72, p < .001, \phi = 0.25$ (DS-1-RS); $\chi^2(1) = 16.95, p < .001, \phi = 0.22$ (DS-2-RS), and days until remediation ($t(169.97) = -4.71, p < .001, d = -0.62$ (DS-1-RT); $t(183.65) = -4.59, p < .001, d = -0.58$ (DS-2-RT)), with *mismatching fingerprints* being remediated more often and in a shorter period of time. This supports that *missing DNSSEC* was eventually more complex to remediate, as it requires more configuration and coordination effort (see Section 2). However, we have to consider that around 46% of websites in the *missing DNSSC* sample are hosted at a DNS provider

14. Link to STT hosted at U2: https://ssh-security-check.aifb.kit.edu. The code is available in [52].

15. See Section 6.3 for an explanation.

TABLE 1: Statistical analysis of the different datasets according to our research questions and the statistical tests used.

	Time to Remediation										Status of Remediation							
	DS-1-RT					DS-2-RT					DS-1-RS				DS-2-RS			
	stat.	df1	df2	p	η^2	stat.	df1	df2	p	η^2	stat.	df	p	φ	stat.	df	p	φ
sender	2.79	1	300	.096	-	2.17	1	335.71	.142	-	1.85	1	.174	-	1.94	1	.164	-
tool vs. no tool	3.34	1	240.05	.069	-	3.76	1	252.52	.053	-	1.77	1	.184	-	1.71	1	.190	-
tool detail	3.42	2	198.85	.035	**.025**	3.62	2	222.37	.028	**.023**	6.25	2	.044	**.14**	5.49	2	.064	-
tool x sender	2.07	5	296	.070	-	2.18	5	332	.056	-	-	-	-	-	-	-	-	-



(a) For tool support (DS-1-RT) (b) For tool support (DS-2-RT)

Figure 4: Kaplan Meier curves for tool support

(Hetzner) that does not offer DNSSEC as confirmed by their customer support, which introduces a huge sample bias.

RQ 1: Tool Support. While we could neither identify significant effects between the two senders (U2 vs. U1), or when considering tool support in general (tool vs. text_only), we found significant differences when looking at the groups in more detail (text_only vs. tool_same vs. tool_diff, see Table 1 for the statistics). However, all tests had only small effect sizes: η^2 = .025 (DS-1-RT), η^2 = .023 (DS-2-RT), φ = .14 (DS-1-RS). Post-hoc analysis showed that significant differences exist between the groups text_only and tool_diff p = .032 (DS-1-RT), p = .025 (DS-2-RT), and more specifically between text_only U2 and tool_diff U1, p = .026 (DS-1-RT), p = .022 (DS-2-RT) (see Table 3 in Appendix D for all results), and applying the Holm-Bonferroni correction confirmed these findings. Survival Analysis confirmed that significant differences exist between the survival distributions for text_only, tool_same, and tool_diff, $\chi^2(2) = 7.68, p = .022$ (DS-1-RT) and $\chi^2(2) = 8.34, p = .015$ (DS-2-RT) (see Figure 4 for the survival functions). Thus, tool support seems to have no effect in general when we compare our text_only condition to both groups that received the link to the STT. However, we see that the tool_diff group (the group where the sender of the notification was different to the institution hosting the STT) has a significantly higher remediation than the text_only group. Specifically significant differences exist between text_only U2 and tool_diff U1.

RQ 2: Interaction Effects. For our datasets, neither the main model nor the main effect for the sender or the interaction between the sender and tool support was significant. We only saw a significance in tool support (text_only vs. tool_same vs. tool_diff). Thus, we cannot prove that remediation is affected by whether the self-test tool is hosted at the same or a different institution. Nevertheless, we can confirm that text_only U2 vs. tool_diff U1 differ significantly with respect to the time to remediation, which is not affected by any interaction effect between sender and tool support.

5.2. Results of the Thematic Analysis

As we only coded unsolicited answers, we could not apply all codes from our codebook. We mainly coded the reasons for the lack of remediation, approaches, challenges, tone, and content. We further considered "incorrect technical implementation" as a separate theme, not a code for the theme "Reasons for lack of remediation". We analyzed in total $n = 50$ email responses (16 to U1 and 16 to U2, totaling 32 for our initial email; 10 to U1 and 8 to U2, totaling 18 for our reminder email). Appendix F.1, Table 4 provides a complete overview of all codes and code frequencies.

Incorrect Technical Implementation. With respect to the theme "incorrect technical implementation", most of the respondents confirmed that either fingerprints did not match (18 responses) or DNSSEC was missing (15 responses). Nine domain owners explained that the implementation is correct for their use case (although the analysis tool reports a misconfiguration), and three domain owners explained that the entries are outdated and no longer used. Two reported incomplete deletion (they forgot to delete old, unused DNS entries), and two domain owners could not identify the problem. One response only stated that our notification was internally forwarded without commenting on the incorrect implementation.

Reasons for Lack of Remediation. Within this category, most responses provided several reasons for lack of remediation. We coded lack of awareness in 16 responses stating that the problem is not an issue for the respective domain owners. Lack of awareness for our study meant that for some of the domain owners the misconfiguration was not seen as a critical security issue. Deliberate lack of maintenance was coded for nine responses. Domain owners explained that the website is not used anymore (three responses), the website has no priority (two responses), the DNS entries were outdated (one response), DNSSEC seems problematic and not useful (one response) or has no priority (one response), or

the misconfiguration is probably not a problem since SSH server authentication via password is deactivated in general (one response).

Finally, for one response we additionally coded reliance on others as the domain owner stated that they had adopted the default configuration in good faith that this was correct. Most interestingly, we also found that seven domain owners had no chance to solve the issue since the problem lies with their external service provider.

Approaches and Challenges. We could see that as a results of our notification most of the domain owners who responded changed (eight responses) their configuration, delegated the implementation of changes (four responses), or planned to change their configuration (ten responses). Seven reported that they do not have the possibility to make changes as DNSSEC is not supported by their hosting provider. Only less than half of the responses (19 responses) mention challenges, mainly lack of time (eight responses) or lack of care, i.e., the domain is not part of the daily business (one response), as well as dependencies or slow processes in organizations, i.e., other entities were needed for remediation (four responses), or the coordination with other stakeholders is complex (one response). Two mentioned that they forgot to delete old DNS entries, which was also coded as a lack of resources, and one mentioned that they are not sure if the old domains might be needed again so they did not change the configuration. Furthermore, two domain owners reported problems within their configuration: One was not able to reproduce the problem, and one mentioned that a wrong automatism created the DNS entries, which had never been updated since. We could not identify lack of technical knowledge as a frequently named challenge as it was only mentioned once.

Sentiment Analysis. The responses we received were overwhelmingly friendly (23 responses) or thankful (13 responses), and the domain owners were keen to provide further information or explanations (23 responses), or wanted to learn more about our research, e.g., about our methodology, motivation, or threat model (ten responses). Two responses provided suggestions for improvements: One recommended writing different notifications for the two scenarios we described (*mismatching fingerprints* and *missing DNSSEC*). The other recommended reporting (sub)domains grouped by CNAMEs in order to reduce the number of subdomains reported to the same domain owner.

We received eight responses that were rather neutral, even if the domain owner did not agree with our findings (two responses). Three domain owners wrote explanations, two gave feedback to our notification, and one asked a question. We only identified four responses in which the domain owners were rather annoyed about our notification. In three of these four responses, the domain owners disagreed with our findings; in one of the four responses, the domain owner raised questions about our motivation and methodology implying disagreement while not directly mentioning it.

6. Discussion

6.1. RQ 1 & RQ 2: Tool Support

Cetin et al. [25] have shown that tool support has no effect compared to text-only notifications, and Maass et al. [18], [19] have shown that providing tool support did not in general increase remediation significantly. However, remediation was successful for more than 90% of the recipients who used the tool [18]. Thus, providing a self-test tool seems recommended [16]–[19], [25], [43]. Therefore, we used parametric (ANOVA) and non-parametric (χ^2-test) methods to see whether linking to a website with a STT can increase remediation significantly. Interestingly, our results were not as clear as in previous studies. While there were no significant effects for the `text_only` vs. `tool` condition, we found that the `text_only` condition from U2 was significantly less effective than the notification sent from U1 with a link to the tool hosted at U2.

We tested interaction effects between the sender and the institution providing the tool, to check if we could find an explanation for the difference. We could not identify interaction effects, but prove that only the main effect for tool support is significant. However, the effect size is very small, so we came to the conclusion that tool support might have an effect when the tool is hosted by an institution that is different than the sender. Furthermore, we found it interesting that the combination TU Berlin as sender, which by name can be associated with a university, and tool linked to the Karlsruhe Institute of Technology, which is not immediately identifiable as university, was most effective.

We have to leave it to future work to investigate the effect of tool support in more detail. We assume that differences could be detected with a larger sample, as our analysis only identified small effect sizes < 0.1, while we calculated our sample size based on medium effect sizes. Since in our study the `tool_diff` condition was most effective, it would especially be interesting to analyze if differences are more obvious when sender and hoster have different affiliations, and the institution hosting the tool is not clearly identifiable as a university (e.g., university sender, but STT provided by a company). However, until further results are available, it does not seem to be a huge disadvantage if no tool support can be provided for a specific notification campaign.

6.2. RQ 3 & RQ 4: Responses and Remediation Behavior

While we measured remediation as the correction of the misconfigured SSHFP records and activation of DNSSEC, we learned from the response and our analysis that not all domain owners remediated as we anticipated. Instead, some domain owners removed the (incorrect) SSHFP records of affected (sub)domains. While, e.g., removal of subdomains made the STT stop reporting an issue (since no SSHFP records could be analyzed), it leads to users facing SSH's

default behavior of manual HKV, which may pose some risks as described in section 2. However, we can assume that action was taken based on our notification, making our notification campaign effective, although the misconfiguration was not fully addressed.

Further, we learned from the response emails we got that our remediation rate was also influenced by individual configurations the domain owners made. From the responses, 16 out of 50 domain owners explicitly stated that they would not make any changes because their configuration is otherwise secure or made on purpose (i.e., multiple SSH servers on different ports), which leads to lower remediation rates although our notification was read and processed – meaning the goal of our notification campaign, raising awareness for possibly insecure SSHFP configurations, was met.

We also learned that some domain owners did not even have a chance to remediate by enabling DNSSEC, because their DNS hosting provider has not implemented the possibility to use DNSSEC at all. We approached this particular provider with our findings, but they informed us that the implementation is not planned in the near future, because enabling DNSSEC for their customers would take a lot of time and personnel. At the same time, demand is, at least in Germany, too low for this to be planned at the moment. However, they will monitor demand for this feature and the general market situation on an ongoing basis.

Overall, we saw that domain owners drew very distinct decisions based on our notification. This also means that while we planned to measure the effectiveness of our notification campaign by analyzing remediation based on our scan results, like in previous notification studies, we found that this method does not represent our findings properly. Non-remediation, even if the domain owners had valid reasons not to do so, would imply that our notifications were ineffective when, in fact, we successfully reached the domain owners.

As already discussed in Cetin et al. [25], a low remediation rate might have several reasons researchers don't know about, and only the interpretation of scan results might not sufficiently reflect the effectiveness of a notification campaign. Li et al. [17] or Zeng et al. [24], e.g., found that domain owners draw decisions based on individual cost-benefit analysis, which results in leaving some services vulnerable or disagreeing with the researcher's risk assessment. Durumeric et al. [14] found that missing possibilities to remediate, e.g., no server access, were reasons for low remediation rates. Further, Stöver et al. [44] found that within the notification campaign of [19], ambiguous responsibilities in organizations impacted the lack of remediation, while lack of resources (e.g., time) or slow organizational processes delayed or hindered remediation.

Thus, we argue that non-remediation does not automatically mean the notification was not received, not processed, or the recipients lack awareness or technical knowledge. Instead, it is very likely that in some cases the notifications are processed differently than expected by the researchers, which, in turn, does not make the notification campaign ineffective. This could also be the reason why previous notification experiments have rarely found significant effects with respect to the sender, tool support, or framings – even if surveys with website owners have shown that this should have an effect [14], [17], [19], [20], [24], [25]. We recommend that future notification studies should pay more attention to this, and always consider individual decisions when reporting remediation rates, e.g., by including qualitative research. Furthermore, identifying reasons for non-remediation at a larger scale could help derive patterns and create customized notifications for different user groups.

6.3. Limitations and Future Work

Internal Validity. Several factors might influence the internal validity of our results. We observed delivery issues, particularly with the generic `hostmaster@<domain>` email addresses (118 out of 214 bounced), which primarily impacted the sample size. Reasons for delivery failures for manually collected email addresses (4.6%) can be human errors while retrieving the information, poorly maintained contact information, network issues, or misconfiguration of receiving mail servers. We see that these numbers are in line with related work that also retrieved contact addresses manually [18], [19].

During our analysis, we noticed that several (sub-) domains had CNAME records configured, which delegated the resolution of SSHFP entries to a referenced domain. This is not an issue if the second-level domains are identical to the CNAME records (similar to the example in Figure 2 in Appendix A), since the domain owner remains in control. However, if the referenced second-level domain is different from the analyzed one, the notified domain owners can only remediate the misconfiguration by removing the subdomain or contacting the operator of the referenced domain, as they cannot change records of a domain not under their control. We addressed this issue by excluding those (sub-)domains from our analysis, where we had definite proof that one domain owner took care of several domains in our sample to avoid biases in our analysis.

Besides CNAMEs, further (sub-) domains might be operated by or related to several domain owners we notified. Those domains where we had clear indicators that they were connected to each other or operated by one and the same domain owner were assigned to the same group before we sent our notifications. However, we cannot be entirely sure that domain owners received notifications for more than one group if a domain owner operates several (sub-) domains but does not appear as a contact in the imprint.

One factor limiting the amount of results was network timeouts, caused by the way the analysis tool of Neef et al. [13] retrieves the server-side SSH public keys. We observed that for certain domains, which all resolved to the same IP address, no connection to the corresponding SSH server could be established after a certain number of initially successful connections. We believe that these systems limited the number of connections originating from a single IP address at some point. While such network-based mitigations are out of our control, we reduced the

impact by slowing down the scanning rate and running the scan multiple times per day. Furthermore, we subsequently excluded domains with periodic scan errors for more than four consecutive data points (i.e., a whole day) from our analysis to prohibit bias in our survival analysis.

External Validity. Our domain sample is exclusively focused on German domains, risking that the generalization of our results is limited in terms of language and cultural diversity. However, we deliberately decided against extending our study to other countries. According to Maass et al. [19] tailoring notifications to only one country has the advantage that notifications are better comprehended, and trust in the notification is higher as names of organizations involved are better recognized. Additionally, we would have encountered difficulties in obtaining accurate contact information, as other countries may not require domain owners to provide contact details in an imprint.

Due to domains being excluded, we did not reach the calculated sample size for the single-factor ANOVA for DS-1-RT, the two-factor ANOVA, and the Survival Analysis. This introduces the risk of the affected tests being underpowered, potentially leading to a type II error. Please note that besides the single-factor ANOVA for DS-1-RT, the two-factor ANOVA, and the Survival Analysis, all other tests had a sufficient sample size. Furthermore, both, single factor ANOVA for DS-1-RT and Survival Analysis, also showed significant results for tool support, indicating that our tests were able to detect significant differences. Thus, we are confident that the reduced sample size did not affect the adequacy of our statistical analysis.

Lastly, it might be that the two senders are not as distinguishable as we assumed. Recipients unfamiliar with the universities may not have associated the same differences with the names, but instead perceived both senders as originating from the same domain. This could have weakened the interaction effect between tool support and sender. Investigating interaction effect in future work seems promising, as, we identified weak significant differences between `text_only` U2 vs. `tool_diff` U1.

We recommend addressing the limitations mentioned above in future work by increasing the sample size, extending our study design to other countries, and test the interaction effect between sender and tool support for different senders – as described in Section 6 – to see if similar results can be observed.

We also acknowledge that the misconfiguration we identified is rather niche, and our sample likely consists mainly of tech-savvy domain owners. We could not derive any limitations to our study results from this, but future work will need to take this into account when comparing to our work.

7. Conclusion

By conducting a 2×3 randomized controlled notification experiment, our study is the first that aims to raise awareness among domain owners about the risks and existence of SSHFP misconfigurations, specifically mismatching host key fingerprints between the DNS and the SSH server, and missing DNSSEC, that were found prevalent in [12], [13]. Within our notification campaign, we notified 518 domain owners via email and measured the effect of sender and tool support to complement previous findings by adding further evidence if – and if so how – tool support has an impact on remediation. Most importantly, our study is the first to investigate possible interaction effects between the sender of the notification and the institution hosting the tool.

With respect to our research questions, we found that the effect of tool support needs further investigation. While we could not identify a significant effect of tool support vs. text-only notifications, we found that text-only notifications sent out by U2 were significantly less effective than notifications sent out by U1 with a link to the tool hosted at U2. This raises the question of whether a combination in which the sender of the notification and the institution hosting the tool are clearly different (e.g., sender with a university affiliation and tool support provided by the national CERT or the hosting provider) increases the effects we found.

While only around 30.0% of domain owners remediated the misconfiguration after our notification, we learned from the thematic analysis of the responses we received that low remediation rates do not automatically represent a lack of awareness or lack of technical knowledge among the domain owners. Instead, we found that our notification triggered several responses in which domain owners explained their situation, leading us to conclude that a low remediation rate does not necessarily make a notification campaign ineffective if raising awareness for a certain issue is the main priority. Future work should take into account that domain owners might apply individual cost-benefit analyses or face different challenges, like a missing server access or a contractor who does not provide the required infrastructure. Thus, results of notification campaigns should be discussed accordingly, and preferably always take qualitative data like responses from domain owners, surveys or interviews into account.

Code and Data

All code and data (as far as permissible by GDPR) are made publicly available following the FAIR principles to allow reproduction. The repository with code (e.g., analysis tool) and data (e.g., website data for the STT) is available at https://github.com/gehaxelt/SSHFP-Notification-Study-AE.

Acknowledgements

This research is supported by the German Federal Ministry of Education and Research as part of the INSPECTION project (Zuwendungsnummer 16KIS1113), and by funding from the topic Engineering Secure Systems, topic 46.23.01 Methods for Engineering Secure Systems, of the Helmholtz Association (HGF) and by KASTEL Security Research Labs.

References

[1] T. Lehtinen, "RFC 4250: The Secure Shell (SSH) Protocol Assigned Numbers," 2006. [Online]. Available: https://www.rfc-editor.org/rfc/rfc4250

[2] T. Ylonen, "RFC 4251: The Secure Shell (SSH) Protocol Architecture," 2006. [Online]. Available: https://www.rfc-editor.org/rfc/rfc4251

[3] ——, "RFC 4252: The Secure Shell (SSH) Authentication Protocol," 2006. [Online]. Available: https://www.rfc-editor.org/rfc/rfc4252

[4] ——, "RFC 4253: The Secure Shell (SSH) Transport Layer Protocol," 2006. [Online]. Available: https://www.rfc-editor.org/rfc/rfc4253

[5] ——, "RFC 4254: The Secure Shell (SSH) Connection Protocol," 2006. [Online]. Available: https://www.rfc-editor.org/rfc/rfc4254

[6] J. Schlyter, "RFC 4255: Using DNS to Securely Publish Secure Shell (SSH) Key Fingerprints," 2006. [Online]. Available: https://www.rfc-editor.org/rfc/rfc4255

[7] F. Cusack, "RFC 4256: Generic Message Exchange Authentication for the Secure Shell Protocol (SSH)," 2006. [Online]. Available: https://www.rfc-editor.org/rfc/rfc4256

[8] P. Gutmann, "Do Users Verify SSH Keys?" 2011. [Online]. Available: https://www.usenix.org/system/files/login/articles/105484-Gutmann.pdf

[9] M. Kaier, "ssh-mitm/ssh-mitm: SSH-MITM - ssh audits made simple," 2007. [Online]. Available: https://github.com/ssh-mitm/ssh-mitm

[10] M. Oosterhof, "Cowrie — Cowrie 2.5.0 documentation," 2009. [Online]. Available: https://cowrie.readthedocs.io/en/latest/README.html

[11] F. Valsorda, "FiloSottile/whoami.filippo.io: A ssh server that knows who you are. $ ssh whoami.filippo.io," 2015. [Online]. Available: https://github.com/FiloSottile/whoami.filippo.io

[12] O. Gasser, R. Holz, and G. Carle, "A deeper understanding of SSH: Results from Internet-wide scans," in *2014 IEEE Network Operations and Management Symposium (NOMS)*. IEEE, 2014, pp. 1–9.

[13] S. Neef and N. Wisiol, "Oh SSH-it, What's My Fingerprint? A Large-Scale Analysis of SSH Host Key Fingerprint Verification Records in the DNS," in *International Conference on Cryptology and Network Security*. Springer, 2022, pp. 71–88.

[14] Z. Durumeric, F. Li, J. Kasten, J. Amann, J. Beekman, M. Payer, N. Weaver, D. Adrian, V. Paxson, M. Bailey, and J. A. Halderman, "The Matter of Heartbleed," in *Proceedings of the 2014 Conference on Internet Measurement Conference, IMC '14*, ser. IMC '14. Vancouver, BC, Canada: Association for Computing Machinery, 2014, pp. 475–488. [Online]. Available: https://doi.org/10.1145/2663716.2663755

[15] M. Kührer, T. Hupperich, C. Rossow, and T. Holz, "Exit from Hell? Reducing the Impact of Amplification DDoS Attacks," in *23rd USENIX Security Symposium (USENIX Security 14)*. San Diego, CA: USENIX Association, 2014, pp. 111–125. [Online]. Available: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/kuhrer

[16] F. Li, G. Ho, E. Kuan, Y. Niu, L. Ballard, K. Thomas, E. Bursztein, and V. Paxson, "Remedying Web Hijacking: Notification Effectiveness and Webmaster Comprehension," in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16. Montreal, Quebec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 1009–1019. [Online]. Available: https://doi.org/10.1145/2872427.2883039

[17] F. Li, Z. Durumeric, J. Czyz, M. Karami, M. Bailey, D. McCoy, S. Savage, and V. Paxson, "You've Got Vulnerability: Exploring Effective Vulnerability Notifications," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 1033–1050. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/li

[18] M. Maass, M.-P. Clement, and M. Hollick, "Snail Mail Beats Email Any Day: On Effective Operator Security Notifications in the Internet," in *The 16th International Conference on Availability, Reliability and Security (ARES 2021)*. Vienna, Austria: ACM, New York, NY, USA, 2021, pp. 1–13. [Online]. Available: https://www.readcube.com/library/42cf3704-a798-4336-b319-363ceea244b9:2d81a4f2-09b9-4582-88c0-04a1e1096fac

[19] M. Maass, A. Stöver, H. Pridöhl, S. Bretthauer, D. Herrmann, M. Hollick, and I. Spiecker, "Effective notification campaigns on the web: A matter of Trust, Framing, and Support," in *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 2021, pp. 2489–2506. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/maass

[20] B. Stock, G. Pellegrino, F. Li, M. Backes, and C. Rossow, "Didn't You Hear Me? - Towards More Successful Web Vulnerability Notifications," in *Proceedings of the 25th Annual Symposium on Network and Distributed System Security (NDSS '18)*, 2018, pp. 1 – 15. [Online]. Available: https://swag.cispa.saarland/papers/stock2018notification.pdf

[21] B. Stock, G. Pellegrino, C. Rossow, M. Johns, and M. Backes, "Hey, You Have a Problem: On the Feasibility of Large-Scale Web Vulnerability Notification," in *25th USENIX Security Symposium (USENIX Security 16)*. Austin, TX: USENIX Association, 2016, pp. 1015–1032. [Online]. Available: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/stock

[22] StopBadware and Commtouch, "Compromised Websites: An Owner's Perspective," pp. 1 – 15, 2012. [Online]. Available: https://www.stopbadware.org/files/compromised-websites-an-owners-perspective.pdf

[23] M. Vasek and T. Moore, "Do Malware Reports Expedite Cleanup? An Experimental Study," in *5th Workshop on Cyber Security Experimentation and Test, CSET '12, Bellevue, WA, USA, August 6, 2012*. USENIX Association, 2012, pp. 1 – 8. [Online]. Available: https://www.usenix.org/conference/cset12/workshop-program/presentation/vasek

[24] E. Zeng, F. Li, E. Stark, A. P. Felt, and P. Tabriz, "Fixing HTTPS Misconfigurations at Scale: An Experiment with Security Notifications," in *The 2019 Workshop on the Economics of Information Security (2019)*, Boston, MA, 2019, pp. 1 – 19. [Online]. Available: https://www.semanticscholar.org/paper/Fixing-HTTPS-Misconfigurations-at-Scale%3A-An-with-Zeng-Li/b22c522c6201f8545e1626deaf6ca43db52444d7

[25] F. O. Çetin, C. H. Ganan, M. T. Korczynski, and M. J. G. v. Eeten, "Make notifications great again: learning how to notify in the age of large-scale vulnerability scanning," ser. 16th Workshop on the Economics of Information Security (WEIS 2017), San Diego, 2017, pp. 1–23. [Online]. Available: http://resolver.tudelft.nl/uuid:621f4a4f-e5d9-4f04-abc4-46252f9db3db

[26] O. Çetin, L. Altena, C. Gañán, and M. v. Eeten, "Let Me Out! Evaluating the Effectiveness of Quarantining Compromised Users in Walled Gardens," in *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. Baltimore, MD: USENIX Association, 2018, pp. 251–263. [Online]. Available: https://www.usenix.org/conference/soups2018/presentation/cetin

[27] O. Çetin, C. Gañán, L. Altena, S. Tajalizadehkhoob, and M. v. Eeten, "Tell Me You Fixed It: Evaluating Vulnerability Notifications via Quarantine Network," *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, vol. 00, pp. 326–339, 2019. [Online]. Available: https://www.readcube.com/library/42cf3704-a798-4336-b319-363ceea244b9:55032f97-a10f-4272-baa8-31cb68b24da4

[28] O. Çetin, M. H. Jhaveri, C. Gañán, M. v. Eeten, and T. Moore, "Understanding the role of sender reputation in abuse reporting and cleanup," *Journal of Cybersecurity*, vol. 2, no. 1, pp. 83–98, 2016. [Online]. Available: https://academic.oup.com/cybersecurity/article/2/1/83/2629556

[29] P. Mockapetris, "RFC 882: Domain names: Concepts and facilities," 1983. [Online]. Available: https://www.rfc-editor.org/rfc/rfc882

[30] P. Hoffmann, "RFC 9364: DNS Security Extensions (DNSSEC)," 2023. [Online]. Available: https://www.rfc-editor.org/rfc/rfc9364

[31] G. Schmid, "Thirty Years of DNS Insecurity: Current Issues and Perspectives," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, p. 2429–2459, 2021. [Online]. Available: http://dx.doi.org/10.1109/COMST.2021.3105741

[32] A. Herzberg and H. Shulman, "Towards Adoption of DNSSEC: Availability and Security Challenges," Cryptology ePrint Archive, Paper 2013/254, 2013. [Online]. Available: https://eprint.iacr.org/2013/254

[33] R. Andrews, D. A. Hahn, and A. G. Bardas, "Measuring the Prevalence of the Password Authentication Vulnerability in SSH," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–7.

[34] M. Thankappan, H. Rifà-Pous, and C. Garrigues, "Multi-Channel Man-in-the-Middle attacks against protected Wi-Fi networks: A state of the art review," *Expert Systems with Applications*, vol. 210, p. 118401, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422015093

[35] S. K. Singh, S. Gautam, C. Cartier, S. Patil, and R. Ricci, "Where The Wild Things Are: Brute-Force SSH Attacks In The Wild And How To Stop Them," in *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*. Santa Clara, CA: USENIX Association, Apr. 2024, pp. 1731–1750. [Online]. Available: https://www.usenix.org/conference/nsdi24/presentation/singh-sachin

[36] K. Ryan, K. He, G. A. Sullivan, and N. Heninger, "Passive SSH Key Compromise via Lattices," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 2886–2900. [Online]. Available: https://doi.org/10.1145/3576915.3616629

[37] C. Hetzler, Z. Chen, and T. M. Khan, "Analysis of SSH Honeypot Effectiveness," in *Advances in Information and Communication*, K. Arai, Ed. Cham: Springer Nature Switzerland, 2023, pp. 759–782.

[38] J. Piet, A. Sharma, V. Paxson, and D. Wagner, "Network Detection of Interactive SSH Impostors Using Deep Learning," in *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, Aug. 2023, pp. 4283–4300. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/piet

[39] M. Başer, E. Y. Güven, and M. A. Aydın, "SSH and Telnet Protocols Attack Analysis Using Honeypot Technique: Analysis of SSH AND TELNET Honeypot," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, 2021, pp. 806–811.

[40] M. M. Raikar and S. M. Meena, "SSH brute force attack mitigation in Internet of Things (IoT) network : An edge device security measure," in *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, 2021, pp. 72–77.

[41] D. Sikeridis, P. Kampanakis, and M. Devetsikiotis, "Assessing the overhead of post-quantum cryptography in TLS 1.3 and SSH," in *Proceedings of the 16th International Conference on Emerging Networking EXperiments and Technologies*, ser. CoNEXT '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 149–156. [Online]. Available: https://doi.org/10.1145/3386367.3431305

[42] C. Utz, M. Michels, M. Degeling, N. Marnau, and B. Stock, "Comparing Large-Scale Privacy and Security Notifications," *Proceedings on Privacy Enhancing Technologies*, vol. 2023, no. 3, p. 173–193, 2023.

[43] M. Maaß, H. Pridöhl, D. Herrmann, and M. Hollick, "Best Practices for Notification Studies for Security and Privacy Issues on the Internet," in *The 16th International Conference on Availability, Reliability and Security*, ser. The 16th International Conference on Availability, Reliability and Security. Vienna, Austria: Association for Computing Machinery, 2021, pp. 1–10.

[44] A. Stöver, N. Gerber, H. Pridöhl, M. Maass, S. Bretthauer, I. S. genannt Döhmann, M. Hollick, and D. Herrmann, "How Website Owners Face Privacy Issues: Thematic Analysis of Responses from a Covert Notification Study Reveals Diverse Circumstances and Challenges," *Proc. Priv. Enhancing Technol.*, vol. 2023, pp. 251–264, 2023. [Online]. Available: https://petsymposium.org/popets/2023/popets-2023-0051.php

[45] "Wir sind .de - DENIC eG," 2024. [Online]. Available: https://www.denic.de/

[46] "OpenINTEL: Active DNS Measurement Project," 2024. [Online]. Available: https://openintel.nl/

[47] R. Sommese, R. van Rijswijk-Deij, and M. Jonker, "This Is a Local Domain: On Amassing Country-Code Top-Level Domains from Public Data," *SIGCOMM Comput. Commun. Rev.*, vol. 54, no. 2, p. 2–9, Aug. 2024. [Online]. Available: https://doi.org/10.1145/3687234.3687236

[48] B. VanderSloot, J. Amann, M. Bernhard, Z. Durumeric, M. Bailey, and J. A. Halderman, "Towards a Complete View of the Certificate Ecosystem," in *Proceedings of the 2016 Internet Measurement Conference*, ser. IMC '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 543–549. [Online]. Available: https://doi.org/10.1145/2987443.2987462

[49] "crt.sh — Certificate Search," 2024. [Online]. Available: https://crt.sh/

[50] "Certstream," 2024. [Online]. Available: https://certstream.calidog.io/

[51] "Attack Surface Management and Threat Hunting Solutions — Censys," 2024. [Online]. Available: https://censys.com/

[52] "SSHFP Notification Study Code & Datasets," 2024, link to repository temporarily removed during shepherding to retain anonymity during the single-blind artifact evaluation. [Online]. Available: https://github.com/gehaxelt/SSHFP-Notification-Study-AE

[53] A. Hennig, H. Dietmann, F. Lehr, M. Mutter, M. Volkamer, and P. Mayer, "'Your Cookie Disclaimer is Not in Line with the Ideas of the GDPR. Why?'," in *Human Aspects of Information Security and Assurance (HAISA 2022)*, ser. IFIP Advances in Information and Communication Technology, vol. 658. Cham: Springer, 2022, pp. 218–227.

[54] D. Crocker, "RFC 2142: Mailbox Names for Common Services, Roles and Functions," 1997. [Online]. Available: https://www.rfc-editor.org/rfc/rfc2142

[55] E. Foudil and Y. Shafranovich, "RFC 9116: A File Format to Aid in Security Vulnerability Disclosure," 2022. [Online]. Available: https://www.rfc-editor.org/rfc/rfc9116

[56] T. Poteat and F. Li, "Who you gonna call? an empirical evaluation of website security.txt deployment," in *Proceedings of the 21st ACM Internet Measurement Conference*, ser. IMC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 526–532. [Online]. Available: https://doi.org/10.1145/3487552.3487841

[57] A. Hennig, F. Neusser, A. A. Pawelek, D. Herrmann, and P. Mayer, "Standing out among the daily spam: How to catch website owners' attention by means of vulnerability notifications," in *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3491101.3519847

[58] M. A. Bujang, N. Sa'at, T. M. I. T. A. B. Sidik, and L. C. Joo, "Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data," *Malaysian Journal of Medical Sciences*, vol. 25, no. 4, pp. 122–130, 2018. [Online]. Available: https://doi.org/10.21315/mjms2018.25.4.12

[59] Q. Wu and K. Lu, "On the Feasibility of Stealthily Introducing Vulnerabilities in Open-Source Software via Hypocrite Commits," 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233479632

[60] P. University, "Princeton-Radboud Study on Privacy Law Implementation," 2021. [Online]. Available: https://privacystudy.cs.princeton.edu/

Appendix A.
HKV prompt, SSHFP example, and relationship

Listing 1 shows the HKV prompt by the OpenSSH client upon the first connection to a yet unknown server. Listing 2 displays an example SSHFP DNS record.

Listing 1: Host Key Verifcation prompt shown by the OpenSSH client upon the first connection to a SSH server.

```
The authenticity of host 'ssh.example.com
    (192.168.0.1)' can't be established.
ED25519 key fingerprint is SHA256:Mz3TPMNH6gwbYF/
    rJqcBtvqYhsNfPD+wfpgYkgmqV2o.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/
    no/[fingerprint])?
```

Listing 2: Example of an SSHFP resource record

```
$> dig SSHFP sshfp.example.com +short
4 2 B6D16BEFB9F8B306E18593791C8E8188
    C69E665DF835D0438EF85D832E22DBDB
```

Appendix B.
Self-Test Tool

Figure 5 shows a screenshot of the page with the analysis results that was shown after the domain owner successfully checked their domain in the STT.

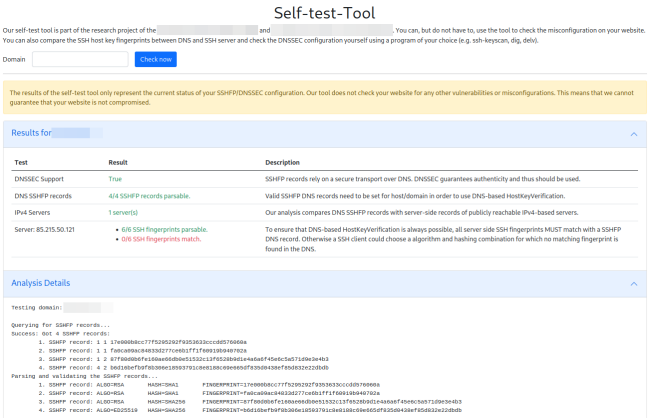


Figure 5: Screenshot of the results shown to the domain owners using the STT.

Appendix C.
Notifications

C.1. Notification text

The notification email template (see Figure 6) was used in combination with Thunderbird's *Mail Merge* extension to fill the placeholders. The template *without link to the STT* (`text_only`) omits the green paragraph about the self-test tool. The `tool_same` email contained a link to the STT hosted at the sender's university, and to the other sender's university in `tool_diff`. For the *reminder notification* we added the text displayed in blue.

```
Greeting {{firstname1}} {{lastname1}},

As part of a research project at {{UNI}}, we are identifying misconfigurations related to your
DNS infrastructure.

We have already contacted you about this on 06.03.2024.

With this message, we merely want to inform you about the behavior we have observed. If you have
already made changes to the SSHFP records, this message likely refers to unconfigured DNSSEC.

Whether you wish to take further action is up to your discretion. As part of our study, you will
not be at any disadvantage if you choose to do nothing further. After this reminder email, we
will not contact you again.

Do you think our email is a mistake? Or are you already aware of the problem? We would greatly
appreciate your feedback in a short reply to this email.

Otherwise, here is the information from our first email again:

# What is the issue?
You are using SSHFP DNS records on the following domains to utilize DNS-based HostKey
verification for SSH:

{{subdomain_1}}
{{subdomain_2}}
{{subdomain_3}}
{{subdomain_4}}

In the course of our investigation, we have found that the entries you have set are not
correctly configured. There are two possible scenarios:

1) The HostKey fingerprints in the SSHFP DNS records do not match the actual HostKeys or HostKey
fingerprints of the SSH server.

2) The SSHFP DNS records are not being securely transmitted because DNSSEC is not configured.

# What can happen?

For scenario 1: If the HostKey fingerprints in the DNS and the corresponding SSH server do not
match, the HostKey verification cannot be automatically performed by the SSH program. This means
that the SSH program (e.g., OpenSSH) will prompt the user for manual verification. Manual
verification carries the risk that it may be done incorrectly or not at all.

For scenario 2: DNS is a plaintext protocol that, without DNSSEC, does not provide guarantees
for the authenticity of the transmitted data. The SSHFP standard requires that these DNS records
be securely transmitted (e.g., via DNSSEC). If this is not the case, network-based attackers can
manipulate or remove the transmitted SSHFP records. This means that automated HostKey
verification will fail or may accept manipulated HostKeys from the attacker.

In both scenarios, this could potentially lead to attackers gaining access to your SSH
credentials and thus to your infrastructure.

You can verify the configuration we identified using a self-test tool. There you will also find
more information on how to proceed. To use the tool, enter {{link}} in your browser and then
verify yourself with the token {{token}}. You can also use the direct link: {{link}}?t={{token}}

We recommend that you resolve the issue. You can consult your IT administrator or DNS hosting
provider on how to set the configuration correctly.

Best regards,
{{Sender}}

--

{{Sender's footer}}
```

Figure 6: Notification and reminder *(blue)* notification text with *(green)* and without link to STT

C.2. Notification emails

Table 2 shows the sent out emails and bounces per sender and group.

Appendix D.
Post-hoc analysis tool support

Table 3 shows the post-hoc analysis for differences between the groups `text_only`, `tool_same`, and `tool_diff`.

Appendix E.
Domain Sampling

Figure 7 visualizes the domain sampling process.

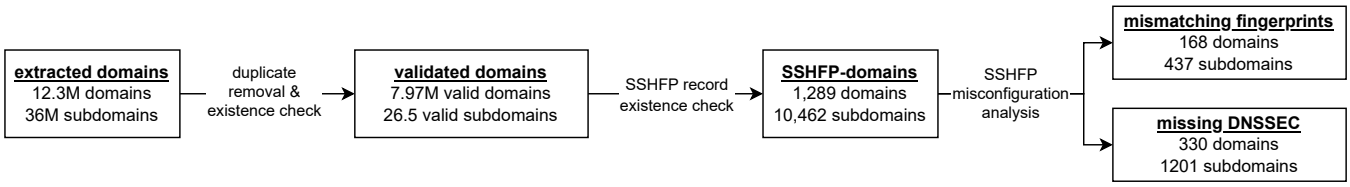


Figure 7: Statistics on analyzed (sub)domains and SSHFP misconfigurations.

TABLE 2: Overview of notification emails sent, bounced, delivered and responses received.

Sender	Group	Emails	Bounces	Delivered	Responses
U1	text_only	87	26	61	7
	tool_same	87	21	66	9
	tool_diff	85	16	69	11†
∑		259	63	196	27
U2	text_only	87	24	63	6
	tool_same	87	22	65	9†
	tool_diff	85	23	62	10
∑		259	69	190	25
∑		518	132	386	52

† This also includes one response that was deleted from the analysis.

TABLE 3: Post-hoc Analysis for Differences between the Groups `text_only`, `tool_same`, and `tool_diff`.

I (Sender)	J (Sender)	Mean diff. (I-J)	Std. err.	Sig.	Sig. (corr)	95% CI Min	95% CI Max
U1	U2	-14.939	10.559	.718	1	-45.65	15.77
	U1_same	-2.911	11.810	1	1	37.24	31.41
	U2_same	-5.018	11.930	.998	1	-39.70	29.66
	U1_diff	19.263	12.229	.617	1	-16.22	54.75
	U2_diff	6.450	13.004	.996	1	-31.46	44.36
U2	U1	14.939	10.559	.718	1	-15.77	45.65
	U1_same	12.028	10.419	.857	1	-18.30	42.36
	U2_same	9.922	10.554	.935	1	-20.82	40.66
	U1_diff	34.203[1]	10.891	.26	.041	2.56	65.84
	U2_diff	21.389	11.755	.460	1	-13.03	55.81
U1_same	U1	2.911	11.810	1	1	-31.41	37.24
	U2	-12.028	10.419	.857	1	-42.36	18.30
	U2_same	-2.107	11.807	1	1	-36.45	32.24
	U1_diff	22.174	12.109	.450	.837	-12.98	57.33
	U2_diff	9.361	12.891	.978	1	-28.24	46.96
U2_same	U1	5.018	11.930	.998	1	-29.66	39.70
	U2	-9.922	10.554	.935	1	-40.66	20.82
	U1_same	2.107	11.807	1	1	-32.24	36.45
	U1_diff	24.281	12.225	.357	.561	-11.22	59.78
	U2_diff	11.467	13.001	.950	1	-26.45	49.39

[1] The mean difference is significant at the 0.05 level.

Appendix F. Thematic Analysis

F.1. Frequency of codes applied

Table 4 shows the observed frequencies of the different codes.

TABLE 4: Frequency of codes applied in the thematic analysis of the email responses.

Variable[16]	**Code: Frequency**
Background[17]	
ownership	professional: 2, private: 1, volunteer: 3
involvement maintenance	operates website with support: 1
motivation	privacy/compliance/security is important: 4
Implementation	
incorrect technical implementation	incorrect implementation: 18, incomplete implementation: 13, implementation is for their use case correct: 9, other: 3
Lack of Remediation	
lack of awareness	not an issue: 16
problem outside of influence	problem lies with external service provider: 5, other: 2
reliance on others' judgements	default settings taken: 1
deliberate lack of maintenance	website has no priority: 2, other: 21
Approaches	
implementation	implement changes themselves: 8, delegate implementation of changes: 4, do not implement changes: 16, other: 13
kind of support	technical support: 2
support instance	study organizer: 2
Challenges	
lack of resources	personal reasons: 8, website is not daily business: 1, other: 3
lack of technical knowledge	lack of technical knowledge: 1
problems with code	error cannot be found: 1, other: 1
dependencies and slow processes in organizations	complex coordination with other stakeholders: 1, other entity needed for fix: 4
Sentiment Analysis	
tone	(rather) neutral: 8, (rather) thankful: 13, (rather) friendly: 23, (rather) annoyed: 4
content	explanation: 23, dissent: 5, question: 10, feedback: 4, expressing thank: 6

F.2. Codebook

Table 5 shows the codebook used for the thematic analysis.

16. For reasons of space we just name the variable in this table. For more explanation on the definition of the variables and the codes see Table 5. Note that we do not report variables here that we did not code.

17. Please note that we did not report on this in the results since we only identified information on background for six responses. We assume from the context of the responses that the majority of the domains in our sample are run in a private context. However, we could not identify clear indicators for that.

TABLE 5: Codebook with themes, categories, explanation and codes that were used for analysis.

Category (variable)	**Explanation**	**Codes**
Theme: Background		
ownership (B_1)	Context of owning the respective domain	professional / private / volunteer / other
involvement development (B_2)	Involvement of the domain owner in the development of the website	did not build website / built website with support / built website without support / other
involvement maintenance (B_3)	Involvement of the domain owner in the maintenance of the domain	has domain operated / operates domain with support / runs domain without support / other
motivation (B_4)	The motivation of the domain owner to tackle the problem	does not care about privacy/security / privacy/security has no high priority / privacy/compliance/security is important / other
Theme: Implementation		
incorrect technical implementation (R_1)	An implementation was incorrect which causes the problem	incorrect implementation / incomplete implementation / wrong code applied by mistake / implementation is for their use case correct / other
Theme: Lack of Remediation		
lack of awareness (R_2)	The domain owner lacks awareness of the problem, either because the domain owner thinks there is no privacy/security issue or the domain owner thinks the problem is not existent	not an issue / already implemented / other
ambiguous responsibilities (R_3)	The responsibilities for the domain or the problem are not clear, so nobody took care of the problem	not aware of responsibility / responsible person not available / other
problem outside of influence (R_4)	The problem lies outside the influence of the domain owner	problem lies with third-party provider / problem lies with external service provider / other
reliance on other's judgments (R_5)	The domain owner relied on the expertise of others	relied on certification / relied on information of external service provider / relied on information of friends / default settings taken / other
deliberate lack of maintenance (R_6)	The domain has deliberately not been maintained	domain has not been maintained for a long time / domain was set up a long time ago / domain has no priority / domain is currently being revised or new domain is being created / other
Theme: Approaches		
implementation (A_1)	Did the domain owner make the changes themselves or with the help of an external?	implement changes themselves / implement changes with support / delegate implementation of changes / do not implement changes / other
kind of support (A_1-1)	[Only if A_1 "implement changes with support" or "delegate implementation of change"] If the domain owner seeks support what kind of support is this?	legal support / technical support / other
support instance (A_1-2)	[Only if A_1 "implement changes with support" or "delegate implementation of change"] What kind of instance does the named support belong to?	study organizer / professional IT person / private IT person / domain provider / hosting provider / other
Theme: Challenges		
lack of resources (C_1)	The lack of resources delays or hinders the remediation of the problem	personal reasons / domain is not daily business / no money for professional help / other
lack of technical knowledge (C_2)	The lack of technical knowledge delays or hinders the remediation of the problem	difficult to keep up to date / lack of technical knowledge / other
problems with code (C_3)	Technical problems exist that delay or hinder the remediation of the problem?	trouble solving the problem / error cannot be found / other
dependencies and slow process in organizations (C_4)	There are some dependencies or in general slow processes that slow down the remediation process	responsible person currently not available / domain is not actively maintained / complex coordination with other stakeholders / other entity needed for fix / other
Theme: Sentiment Analysis		
tone (S_1)	The overall tone of the response email. Only use one code, if there seem to be contradicting tones (e.g. friendly and annoyed), decide which one is most prominent.	(rather) neutral / (rather) thankful / (rather) friendly / (rather) annoyed / (rather) angry / hostile / other
content (S_2)	The content of the response email. If several codes apply, choose the most pre-dominant one. If several codes apply equally, separate them with semicolon.	explanation / dissent / question / request for help / feedback / expressing thank / complaint / authenticating sender / other
Theme: Feedback		
feedback_improvement (F_1)	Feedback mentioned in the response to improve further notifications. If several improvements are mentioned, each are separated by semicolon	problem other than described / group several emails / define the specific problem / include a self-service tool / personalize salutation / other
Theme: Other		
other (OTHER)	Other results that are not in the categories yet, but seem interesting for the analysis	open